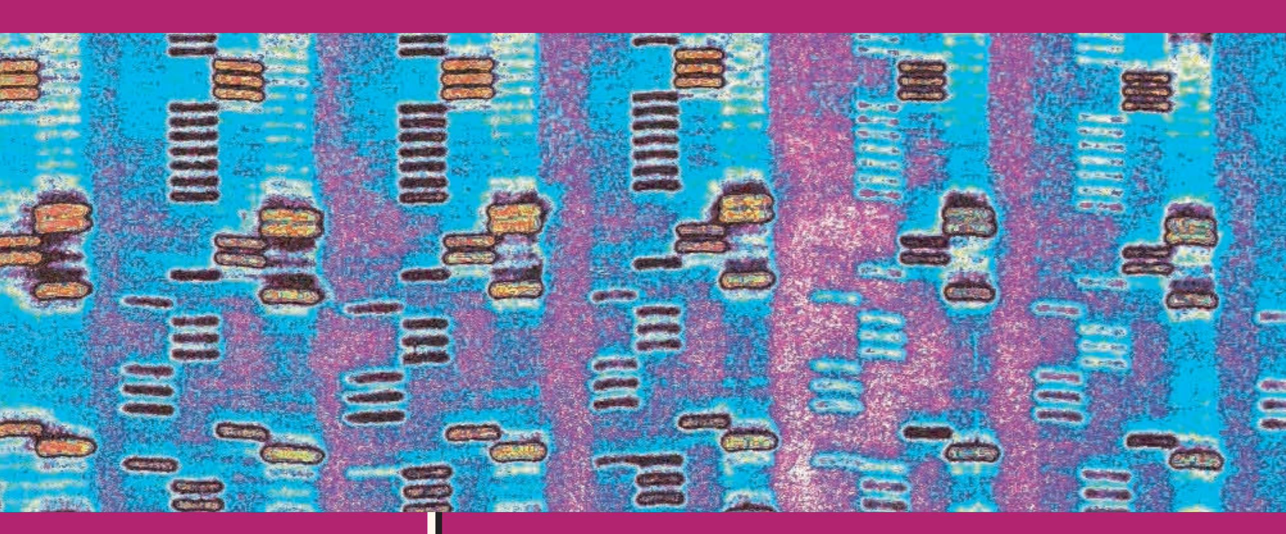


Initiation à la génétique des populations naturelles



Applications aux parasites et à leurs vecteurs

Thierry De Meeûs



Initiation à la génétique des populations naturelles

Application aux parasites et à leurs vecteurs

Initiation à la génétique des populations naturelles

Application aux parasites et à leurs vecteurs

Thierry
De Meeûs

IRD Éditions
INSTITUT DE RECHERCHE
POUR LE DÉVELOPPEMENT

Collection DIDACTIQUES

Marseille, 2012

Préparation éditoriale
Yolande Cavallazzi

Mise en page
Desk (53)

Maquette de couverture
Michelle Saint-Léger

Maquette intérieure
Pierre Lopez – Aline Lugand/Gris Souris

Coordination, fabrication
Catherine Plasse

Photo de couverture :
©IRD/L. Basco — Séquençage d'ADN.
Retouche graphique : Michelle Saint-Léger

Photo page 4 de couverture :
©IRD/S. Ravel — *G. palpalis gambiensis*, accouplement.

©IRD/J.-L. Frézil — *Trypanosoma gambiense* sur frottis de sang.

ISBN : 978-2-7099-1732-2
ISSN : 1142-2580

À Soumeïa et Nicolas,
la plus importante partie de ce que je suis.

Sommaire

Avant-propos

Ce document devrait permettre aux débutants et personnes non familiarisées avec la génétique des populations de pouvoir effectuer leurs propres analyses ou au moins de pouvoir mieux comprendre les conseils des spécialistes. Il a été au départ rédigé dans l'urgence pour les étudiants d'un Master de maladies infectieuses. Il a pour vocation d'être utile en premier lieu à ce type d'étudiants, mais il s'adresse également à un public plus large s'intéressant à la structure génétique des populations naturelles et aux inférences qu'il est possible de faire à partir de marqueurs génétiques variables dans le temps et l'espace. C'est pourquoi tous les retours, commentaires et suggestions susceptibles d'améliorer ce travail et d'en permettre une meilleure compréhension seront hautement appréciés. Les formules mathématiques sont nombreuses dans ce manuel. Leur compréhension sur le bout des doigts n'est pas indispensable. Seule la compréhension des grands principes est requise. Cependant, il est clair que d'arriver à comprendre la plupart de ces formules, dont certaines sont vraiment à la base de la génétique des populations, sera d'un très grand secours pour tous ceux qui souhaitent pouvoir s'affranchir le plus possible des spécialistes et de leurs remarques impatientes, parfois désobligeantes. Je me permettrai d'insister sur le fait qu'il ne faut jamais hésiter à demander conseil à un spécialiste. On ne risque en effet que le désagrément de se faire envoyer promener, ce qui n'est pas mortel. Aider ses collègues et en particulier les étudiants est un devoir sacré des chercheurs. Ceux qui refusent de le comprendre ne méritent à mon sens pas leur salaire. Alors mon adage en la matière est « aucune hésitation ! ».

La plupart des exemples et des propos de ce manuel sont centrés sur des problématiques hôte-parasite-vecteur. Cela vient naturellement de mon expérience en la matière. Il n'en reste pas moins que les méthodes décrites ici sont applicables à tous les êtres vivants, même si d'autres outils sont utilisés ailleurs (en particulier, en bactériologie).

Il me faut également remercier un certain nombre de personnes qui par leurs conseils, les échanges que j'ai pu avoir avec elles ou les coups de pouce qu'elles m'ont donnés m'ont permis d'acquérir les compétences qui sont les miennes aujourd'hui. Je ne remercie pas ici ceux qui m'ont aidé dans d'autres domaines de la biologie des populations non directement reliés aux thématiques développées dans le présent manuel. Je tiens d'abord à remercier Jérôme Goudet de m'avoir mis le pied à l'étrier des *F*-statistiques de Wright, de leurs estimateurs et des tests associés, ainsi que de sa patience lors de mon post-doc à Bangor alors que je le harcelais de questions parfois sans doute un peu débiles. Il me faut également remercier Michel Raymond et

François Rousset pour les échanges parfois animés qui m'ont permis de mieux assimiler les statistiques parfois (souvent) non intuitives associées à la génétique des populations. Les discussions avec Jean-François Guégan et les conseils qu'il a pu me prodiguer m'ont grandement aidé, en particulier pour les modèles de régression. Un grand merci également à Éric Elguero, Benjamin Roche et Marc Choisy pour leurs conseils et astuces toujours utiles. Qu'il me soit permis ici de rendre hommage au regretté Anatoli Teriokhin, parti beaucoup trop tôt. Cette liste de remerciements, où les oublis sont obligatoires, serait particulièrement biaisée sans la présence de Christine Chevillon, grande traductrice de Rousset dans le texte devant l'éternel, et donc sans qui une grande partie de mes publications auraient été amputées de paragraphes particulièrement croustillants, voire n'auraient même pas vu le jour. Je me dois également de remercier les étudiants que j'ai encadrés et dont les remarques, révoltes et questionnements m'ont particulièrement enrichi, et pas seulement en termes de titres et travaux. Je pense plus particulièrement à Franck Prugnolle, mais aussi à Damien Caillaud. Merci aussi à Michel Tibayrenc d'avoir ouvert la voie de l'épidémiologie moléculaire et de m'avoir accueilli dans son laboratoire en 1999 et laissé entière liberté d'y mener mes recherches. Merci à tous mes collaborateurs, chercheurs, étudiants ou post-docs dont la liste exhaustive serait fastidieuse mais dont les principaux, non encore cités ci-dessus sont : Francisco Ayala, François Balloux, Anne-Laure Bañuls, Nicolas Barré, Adrien-Marie-Gaston Belem, Jérémy Bouyer, Bruno Bucheton, Mamadou Camara, Michel de Garine-Wichatitsky, Sylvie Hurtrez-Boussès, Florent Kempf, Mathurin Koffi, Naférima Koné, Laurent Lehmann, Annette MacLeod, Karen D. McCoy, François Nébavi, Flobert Njiokou, Denis Roze, Issa Sidibé, Gustave Simo, André Théron, Sophie Ravel, Virginie Rougeron et j'en oublie surement.

Cependant, cette liste de personnes à remercier souffrirait d'une carence grave sans la présence des chercheurs de l'UMR IRD/Cirad 177 Intertryp qui ont la bonté de tolérer ma présence dans leur équipe. Merci à Gérard Cuny de m'avoir accueilli sans poser de question. Un tsé-tsé grand merci à Philippe Solano, maintenant vieux complice et à l'origine de mon intégration dans mon équipe actuelle et merci à Vincent Jamonneau de m'avoir permis de toucher au monde fascinant des trypanosomes africains. Merci à vous de me permettre de vivre cette expérience enthousiasmante au Burkina Faso. Merci aussi à tout le personnel du Cirdes et à mes étudiants burkinabè Jacques Kaboré et Modou Séré et merci à tous les étudiants ayant suivi (ou subi) mes cours et qui par leurs questions m'ont permis d'améliorer la vision que j'ai de mon travail.

Merci à Tatiana Giraud (TG) d'avoir accepté le travail ingrat et combien fastidieux de relire ce travail et d'avoir ainsi contribué à une bien meilleure lisibilité de ce manuel.

Merci à toute l'équipe des Editions de l'IRD pour leur travail et leur infinie patience, en particulier Yolande Cavallazzi, sans qui un nombre incalculable de coquilles

continueraient à infester ma prose, Catherine Plasse, Michelle Saint-Léger et Thomas Mourier.

Avant de terminer cet avant-propos, et parce que le monde de la recherche peut s'avérer parfois très (trop) compétitif, j'aimerais exprimer quelques opinions personnelles à destination des plus jeunes. La seule compétition qui mérite un intérêt est celle que l'on engage contre soi-même, et les autres, en particulier les collègues, sont là pour nous aider à mener à bien ce combat. Pour vaincre il faut renoncer à gagner. Je remercie donc tous mes échecs de m'avoir rendu meilleur.

Et enfin pour paraphraser un proverbe africain d'origine incertaine « Mais entouka, ce qui est sûr c'est que ça va aller ! ».

Introduction

Les organismes parasites représentent une part significative de la biodiversité répertoriée (espèces décrites) (De Meeûs et Renaud, 2002) et malgré la récente explosion des études moléculaires des populations naturelles, celles concernant les systèmes hôte-parasite sont encore beaucoup trop rares (Criscione *et al.*, 2005). Les agents pathogènes et leurs vecteurs sont en effet des organismes dont la biologie des populations, leur écologie, leur mode de reproduction, déplacements, taille de populations sont difficiles (voire impossibles) d'accès par observation directe. Or, la compréhension de l'épidémiologie d'une maladie infectieuse ou parasitaire, ainsi que l'évaluation des risques d'invasion ou d'épidémie, de même que la perception du risque de diffusion de gènes de résistance ou de l'effet d'une stratégie de lutte sur les populations cibles, ne peuvent se passer d'une connaissance minimale du fonctionnement des populations concernées. Par conséquent, l'écologie, les modalités et/ou stratégies reproductrices (reproduction sexuée ou asexuée, croisements au hasard ou autofécondation partielle ou totale, etc.), la dispersion, la taille des population de parasites et de leurs vecteurs sont des notions clés qui ne peuvent, la plupart du temps n'être inférées que par des méthodes que Slatkin (1985) appelle « indirectes » (Nadler, 1995 ; De Meeûs *et al.*, 2002a, b). Dans ce cas de figure, les méthodes indirectes se caractérisent par l'utilisation de marqueurs moléculaires (génétiques) polymorphes (variables) et l'étude des variations de ces marqueurs dans les individus, entre individus et entre un certain nombre de groupes d'individus prédéfinis comme sous-populations ou plus justement comme sous-échantillons. L'hypothèse de base sous-tendue est que la distribution de la variabilité génétique reflète les paramètres écologiques cités plus haut. Or cette hypothèse, en soi, est assez raisonnable. Nous verrons cependant que d'autres hypothèses plus spécifiques sont souvent requises pour préciser les inférences désirées. L'utilisation de marqueurs génétiques permet d'avoir accès indirectement à des informations clés sur la biologie des populations naturelles des êtres vivants. Comme nous le verrons, ces méthodes s'appliquent également aux organismes non parasites. Les outils de la génétique des populations offrent à cet égard un avantage que des méthodes basées sur l'observation ou la capture des organismes ne donnent pas. L'utilisation de matériel héréditaire (transmissible) ouvre l'accès à des événements rares et passés, par définition peu ou pas accessibles à l'observateur, même au cours de campagnes intensives d'observations de terrain (Prugnolle et De Meeûs, 2002). Ceci ne retire rien aux mérites des méthodes dites directes et, quand cela est possible, l'empiriste aura tout à gagner à utiliser les deux méthodes conjointement sur le même matériel. Cela est

malheureusement encore trop peu souvent mis en œuvre. Les quelques études existantes réalisées soit sur les mêmes individus (Wilson *et al.*, 2004), soit en échantillonnages différés (Hauswaldt et Glenn, 2005 ; Van Bekkum *et al.*, 2006 ; Hoffman *et al.*, 2006) tendent à montrer, par la différence des résultats obtenus, la complémentarité des deux approches ou plus rarement une convergence étonnante (Watts *et al.*, 2007 ; Bouyer *et al.*, 2009 ; De Garine *et al.*, 2009). Cela étant, pour les systèmes hôte-parasite-vecteur, le marquage est le plus souvent impossible de toutes façons (au moins pour le pathogène). Il faut cependant citer ici la tentative méritoire de Chlyeh *et al.* (2002) sur les bulins, hôtes intermédiaires de schistosomes et sur les tsé-tsé sur lesquelles nous reviendrons.

L'accès à ce type d'information n'a pas qu'un intérêt académique, il n'est pas non plus réductible à un simple divertissement intellectuel (Milgroom, 1996 ; Tibayrenc, 1998, 1999 ; Taylor *et al.*, 1999 ; Criscione *et al.*, 2005). « *Population structure and mating system of pathogens are tightly linked biological phenomena with crucial consequences on the epidemiology of transmissible diseases* » (Tibayrenc et Ayala, 2002). Ces informations peuvent en effet s'avérer cruciales pour le contrôle de certaines maladies (Milgroom, 1996) et pour les recherches de nouveaux traitements et de mesures de prévention (Taylor *et al.*, 1999) ainsi que pour des évaluations et prédictions plus efficaces quant à l'évolution de résistances aux drogues, antibiotiques et autres biocides (Tibayrenc, 1999). Les recherches utilisant la génétique des populations d'organismes parasites font partie de ce que Tibayrenc (1998) nomme la génétique épidémiologique ou, d'une manière moins ambiguë, l'épidémiologie moléculaire. L'étude de la génétique des populations des parasites, de leurs vecteurs et hôtes peut, comme je viens de le décrire de façon insistante, donner accès à des informations clés sur leur écologie et potentiels évolutifs, mais ceci n'est rendu possible que grâce à une batterie d'outils d'analyses statistiques en perpétuelle croissance et évolution. Le principal objectif de ce manuel est de décrire la plupart des méthodes disponibles à ce jour, leur mérite, leur puissance ainsi que leur limites, les concepts et hypothèses biologiques de base qui permettent leur mise en œuvre et ce de la façon la plus didactique possible. Pour des revues plus générales et techniques, le lecteur averti pourra se reporter aux excellentes productions de Criscione et Blouin (2005), Criscione *et al.* (2005), Rousset (2004) (et les références contenues dans ces travaux).

Ce manuel est organisé en deux parties. La première partie est elle-même constituée de trois chapitres : le premier chapitre entreprend de décrire très brièvement les différents types de marqueurs les plus utiles pour les études de génétique des populations naturelles ; le deuxième chapitre traite des concepts de base en génétique des populations et des différents outils (paramètres et estimateurs) les plus utiles pour les études empiriques et le troisième chapitre examine les différentes méthodes statistiques associées à ces descripteurs et estimateurs. Enfin, la seconde partie correspond à une mise en application des chapitres précédents à l'aide de plusieurs exemples

réels que nous allons réanalyser ensemble. La plupart des termes techniques sont définis dans un glossaire que les lecteurs trouveront à la fin de ce manuel. Certaines questions théoriques sont traitées à part dans une partie appelée « Réponses aux questions ». Enfin, le nom, utilité, site web de téléchargement et article associé de tous les logiciels utilisés sont listés en annexe (tabl. 1).

Concepts théoriques et statistiques

1 Qu'est-ce qu'un marqueur génétique ?

NOTIONS PRÉLIMINAIRES

Un marqueur génétique est simplement une portion de l'ADN (acide désoxyribonucléique) de l'organisme étudié, ou un sous-produit codé par cet ADN (comme une protéine). L'ADN est la molécule porteuse de l'hérédité chez tous les êtres vivants[1]. Il importe simplement dans notre cas de toujours regarder ce qui se passe sur cette même portion d'ADN chez tous les individus analysés et, dans la mesure du possible, dans plusieurs échantillons (spatialement et/ou temporellement différents). Il est important que cette portion d'ADN reste la même (même localisation dans le génome, à la même place sur le même chromosome) d'un individu à l'autre, d'où le terme locus. Un locus peut correspondre à un gène (codant pour une fonction quelconque), comme c'est le cas pour les loci enzymatiques (ou iso-enzymatiques), mais il peut aussi correspondre à une zone non codante, et donc à priori non fonctionnelle, de l'ADN comme c'est le cas de la plupart des microsatellites. Enfin, il est important de se souvenir qu'un locus, même non codant, peut se trouver dans un intron, c'est-à-dire dans un gène, et peut donc subir des phénomènes sélectifs par sa liaison physique avec les parties traduites du gène. On appelle ce phénomène l'autostop (ou *hitchhiking* en anglais). Cela reste valable pour un locus situé en dehors de tout gène, mais à proximité d'un locus sélectionné ou simplement parce que le régime de reproduction de l'organisme étudié limite ou empêche la recombinaison entre loci. Dans ce qui suit, je vais considérer que l'organisme étudié est diploïde (comme un moustique ou une tique), c'est-à-dire que chaque portion d'ADN (chaque locus) dispose de deux représentants par individu. Plusieurs loci peuvent être considérés. Nous verrons même qu'il est préférable d'analyser les populations naturelles au travers de plusieurs loci de nature identique (microsatellites ou iso-enzymes). Il n'y a pas de limite supérieure au nombre de loci qu'il faut utiliser, mais l'expérience tend à suggérer que cinq est vraiment une limite inférieure qu'il est plus sage d'éviter quand on peut et que sept commence à représenter un bon chiffre. Pour être informatif, un locus doit être variable (on dit polymorphe), c'est-à-dire qu'il présente plusieurs allèles dans le groupe d'individus échantillonnés et génotypés à ce locus. On trouvera un exemple schématique de marqueurs génétiques polymorphes dans la figure 1.

[1] Exception faite des virus à ARN qui ne sont à proprement parler pas de réels êtres vivants bien que faisant partie du monde vivant.

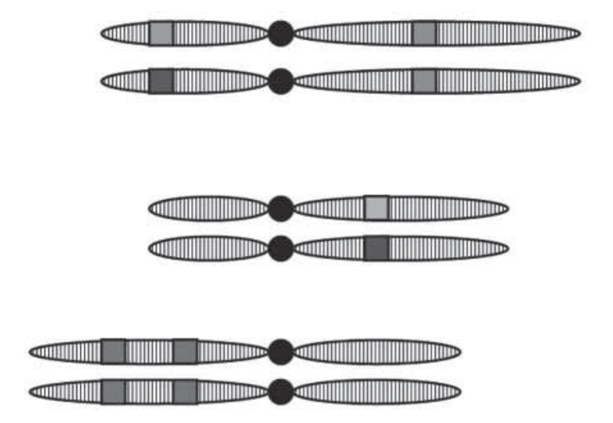


Figure 1
Exemple schématique chez une espèce à trois chromosomes et où cinq marqueurs génétiques (ou loci) ont été définis. On notera que dans cet exemple seuls deux loci sont hétérozygotes (deux allèles différents symbolisés par des couleurs d'intensités différentes) et que les autres sont homozygotes (deux fois le même allèle).

Les mérites et différences entre les différents marqueurs disponibles ont été largement étudiés et ont fait l'objet de nombreuses revues plus ou moins exhaustives que l'on pourra consulter pour plus de précisions (Roderick, 1996 ; Sunnucks, 2000 ; Caterino *et al.*, 2000). Je ne ferai donc qu'effleurer ce sujet que j'ai choisi de subdiviser en trois parties inégales (marqueurs cytoplasmiques, marqueurs nucléaires dominants et marqueurs nucléaires codominants). Nous ne parlerons donc que d'organismes eucaryotes.

MARQUEURS CYTOPLASMIQUES

Les marqueurs cytoplasmiques correspondent à des loci présents dans le génome mitochondrial ou le génome chloroplastique (chez les plantes). Ces marqueurs, et plus particulièrement l'ADN mitochondrial, ont fait l'objet d'un nombre considérable d'études en populations naturelles (Roderick, 1996). L'ADN mitochondrial, ou ADNmt s'est en effet montré extrêmement informatif dans les études phylogéographiques, car il présente des taux d'évolution relativement rapides et ne subit pas de recombinaisons entre loci (Avise *et al.*, 1987 ; Avise, 2000). Cependant, pour les études de génétique des populations, les propriétés de ces marqueurs sont loin d'être idéales et ce pour différentes raisons. Tout d'abord, l'ADNmt présente généralement une hérédité uniparentale, typiquement maternelle bien qu'une transmission paternelle existe chez certains organismes (Le *et al.*, 2002 ; Xu, 2005). La structure

génétique constatée est donc conditionnée par celle observée par un seul des deux sexes chez les organismes dioïques comme le sont de nombreux nématodes, arthropodes et les schistosomes. Par ailleurs, l'effectif efficace (voir encadré 1) pour de tels marqueurs sera toujours difficile à appréhender car dépendant de l'interaction entre divers facteurs tels que le sexe-ratio, le biais de dispersion sexe-spécifique, ainsi que les stratégies de reproduction (PRUGNOLLE et DE MEEÛS, 2002 ; PRUGNOLLE *et al.*, 2003). Ensuite, il est probable que l'ADNmt ne soit pas entièrement neutre (GERBER *et al.*, 2001) et ne serait dans ce cas pas le reflet d'événements démographiques seuls, mais aussi de l'histoire sélective de la population. Enfin, ce sont tous des marqueurs haploïdes qui ne peuvent par conséquent en aucun cas renseigner clairement sur le régime de reproduction local de l'espèce étudiée, au sujet duquel nous verrons que

Encadré 1

L'effectif efficace, noté habituellement N_e, représente une mesure de la vitesse avec laquelle une population de taille N perd de la variabilité génétique par dérive génétique aléatoire. En effet, l'inverse de l'effectif efficace ($1/N_e$ ou $1/2N_e$ pour des diploïdes) donne la probabilité, sur le long terme, que deux allèles d'un même gène (locus) pris au hasard dans la population sont des réplicas (ou des descendants) d'un allèle unique ancestral. Le fait que de tels événements de coalescence interviennent régulièrement (plusieurs gènes descendent alors d'un seul) implique que d'autres allèles doivent avoir disparu. Autrement dit, la diversité génétique s'érode. Le ratio entre l'effectif réel de la population N_c (*census size* qui veut dire taille de recensement en anglais) et l'effectif efficace N_e exprime donc une mesure de la dynamique de quantités associées à la notion de diversité génétique, telle que l'hétéozygotie de la population considérée, par rapport à une population dite idéale. Cette population idéale correspondant en fait à une population qui perdrait sa diversité génétique aussi vite que la population considérée, à la vitesse de $1/N_c$ (ou $1/2N_c$) par génération, de telle sorte que l'effectif efficace de cette population idéale soit égal à l'effectif recensé. Cette caractéristique nécessite une population de taille constante, à générations séparées, hermaphrodite avec rencontre au hasard des gamètes pour former les zygotes et absence de toute forme de sélection, migration ou mutation. À titre d'exemple, considérons une population de bovins de 100 individus composée de 99 ($N_f = 99$) vaches et d'un seul taureau ($N_m = 1$). La taille efficace d'une telle population sera de $N_e = 4N_mN_f/N_c \approx 4$ (voir HARTL et CLARK, 1989 : 86), c'est-à-dire 25 fois plus faible qu'une population de 100 bêtes au sexe-ratio équilibré ($N_f = N_m = 50$). On comprend bien que dans le premier troupeau la diversité génétique s'érode rapidement. D'autres facteurs peuvent influencer l'effritement génétique, parfois en sens inverse comme ce peut être le cas dans les populations subdivisées (ou structurées). Par exemple, dans le cas extrême d'une subdivision totale (pas de transfert de gène entre sous-populations), on atteint une taille efficace infinie, car la diversité génétique se trouve comme gelée au niveau de la population totale même si totalement perdue dans chaque sous-population (chaque sous-population se retrouve rapidement fixée dans un état génétique). Une excellente revue sur le calcul des effectifs efficaces chez les parasites peut être consultée pour ceux qui souhaitent approfondir davantage cette question (CRISCIONE et BLOUIN, 2005).

l'hétérozygotie de marqueurs codominants se montre un auxiliaire précieux. J'ai donc délibérément choisi de ne pas traiter davantage cette famille de marqueurs.

MARQUEURS NUCLÉAIRES DOMINANTS

Avec des marqueurs dominants, les individus hétérozygotes (donc diploïdes) sont vus comme homozygotes pour un des deux allèles présents chez l'individu. Cet allèle est alors appelé dominant par rapport à l'autre allèle qui, invisible à l'état hétérozygote, est qualifié alors de récessif. Ici, le phénotype ne reflète pas fidèlement le génotype. Une des familles les plus connues de marqueurs dominants correspond aux RAPD (*Randomly Amplified Polymorphic DNA*). Des paires d'amorces courtes d'ADN sont utilisées afin d'amplifier par PCR des portions aléatoires d'un ADN cible chaque fois qu'une complémentarité est trouvée. Par conséquent, chez les espèces diploïdes, les individus pour lesquels aucune complémentarité n'existe seront caractérisés par une absence de produit (ADN) amplifié, alors que les individus présentant une séquence complémentaire (hétérozygotes) ou deux (homozygotes pour le complément) présenteront le même produit amplifié, et donc le même phénotype. Il résulte de ceci que seules des fréquences phénotypiques peuvent être estimées avec ce type de marqueurs, alors que les fréquences alléliques demeurent inconnues (à moins de faire des hypothèses très fortes sur la structure des populations). Par ailleurs, la structure génotypique restant elle-même par définition cachée, ainsi en va-t-il des inférences possibles sur le système de reproduction que doit refléter la distribution des allèles dans et entre les individus des mêmes unités de reproduction (sous-échantillons). Qui plus est, et comme déjà mentionné, il est toujours préférable d'étudier plusieurs loci de même nature. Il est impossible de savoir à quoi correspondent les différentes portions d'ADN amplifiées par RAPD de par leur nature aléatoire. On ne peut donc savoir si ces loci sont dans des gènes ou non, quels sont leur taux de mutation, etc. C'est pour ces différentes raisons que les marqueurs dominants en général, et les RAPD en particulier, ne seront pas traités davantage dans ce manuel, car ils sont très loin d'être idéaux pour les analyses de génétique des populations naturelles.

MARQUEURS NUCLÉAIRES CODOMINANTS

Les marqueurs codominants offrent théoriquement l'accès à la structure génotypique complète des individus, c'est-à-dire que tous les génotypes homozygotes et

hétérozygotes sont en principe distinguables. Il existe de nombreuses catégories de marqueurs codominants. Les isoenzymes (ou alloenzymes), les RFLP (*Restriction Fragment Length Polymorphisms*), AFLP (*Amplified Fragment Length Polymorphisms*), microsatellites, minisatellites, MLST (*Multi-Locus Sequence Typing*) et SSCP (*Single-Stranded Conformational Polymorphism*) figurent parmi les plus connus. Les marqueurs SNP (*Single-Nucleotide-Polymorphism*) se montrent extrêmement utiles dans les études d'association, mais ces marqueurs correspondent essentiellement à des loci bi-alléliques (deux allèles seulement), ce qui est loin d'être idéal. De plus, ils présentent des taux de mutations hétérogènes d'un allèle vers l'autre, ce qui est beaucoup plus préjudiciable encore. Il existe en effet un biais clair en faveur des transitions et au détriment des transversions (Vignal *et al.*, 2002). Les SNP ne sont donc pas appropriés aux études de génétique des populations au sens strict du terme. Dans ce qui va suivre je vais surtout traiter des marqueurs isoenzymatiques et microsatellites. Les raisons de cette restriction (si j'ose dire) sont assez simples et pragmatiques. D'abord, ces marqueurs sont les moins chers à mettre en œuvre en travail et moyens (surtout les isoenzymes). De fait, ayant fait moi-même partie d'équipes de recherche françaises avec des moyens modestes (même pour la France, ce qui est tout dire), j'ai participé à ce jour (10-05-2011) à 63 travaux de génétique des populations empiriques (données de terrain) ayant fait l'objet d'une publication dans une revue, dont 17 (~ 30 %) ont utilisé des marqueurs isoenzymatiques. Le reste des études ont utilisé des marqueurs microsatellites qui, en rapport qualité/prix, arrivent juste après les isoenzymes à mon avis. Il en résulte que ce sont les deux types de marqueurs les plus souvent utilisés dans les études de génétique des populations (surtout les microsatellites maintenant car les allozymes sont aujourd'hui plutôt dépassés) en général et surtout ceux que je connais le mieux. Cette dernière raison est sans doute celle qui rend le mieux compte de mon choix qui, de toutes manières, n'a rien de rédhibitoire puisque la presque totalité des informations données dans ce manuel sont applicables à tous les marqueurs codominants. Pour avoir un aperçu des autres techniques, je ne peux qu'encourager le lecteur à consulter les revues existantes (Taylor *et al.*, 1999 ; Caterino *et al.*, 2000 ; Sunnucks, 2000 ; Bougnoux *et al.*, 2004).

Les allozymes

Les allozymes sont en fait des enzymes du métabolisme de base des cellules (comme la Glucose-Phosphate-Isomérase ou GPI qui intervient dans la glycolyse). Pour visualiser de tels marqueurs, les individus ou une partie de leur corps sont broyés dans une solution tampon ou de l'eau distillée et ces extraits sont ensuite déposés soit directement sur gel, soit sur des supports absorbants (comme du papier whatmann) et ces supports absorbants sont eux-mêmes déposés sur ou dans un gel (gel d'amidon, polyacrylamide, acétate de cellulose). Un champ électrique est ensuite appliqué sur le gel. On parle d'électrophorèse des protéines. Les

enzymes étant en général chargées négativement, celles-ci migreront donc vers le pôle positif du champ (anode) et beaucoup plus rarement vers la cathode (si chargées positivement). La vitesse de migration de ces protéines étant fonction de leur charge, la distance parcourue en fin d'électrophorèse reflètera donc aussi cette charge. Les enzymes sont ensuite révélées à l'aide de leur fonction. On utilise en effet le substrat (ou un analogue) qu'elles sont censées transformer, ainsi qu'une substance qui provoque un précipité coloré en présence du produit de la réaction de l'enzyme avec son substrat. À partir de là, plusieurs cas de figure peuvent être rencontrés.

Pas de tache où des traînées non interprétables sont présentes sur le gel

Il faut mettre au point ou passer à un autre locus.

Les taches révélées de tous les individus se retrouvent toutes au même niveau

C'est ce qui se passe, comme dans la figure 2, lorsque la technique ne permet pas de discriminer plusieurs allèles au locus correspondant, soit que ce dernier soit lui-même non variable, soit que les variations existantes ne génèrent pas des allèles aux charges électriques suffisamment différentes pour être perçues par la technique.

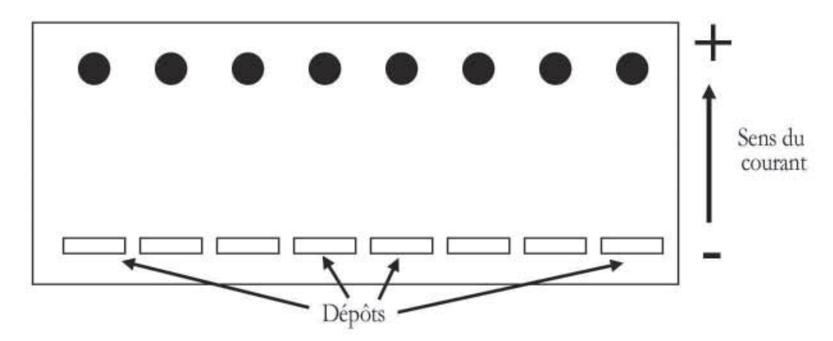


Figure 2
Représentation d'une enzyme monomorphe.

L'enzyme est dite monomorphe, c'est-à-dire que, au moins pour les individus typés (on dit génotypés), il y a absence de polymorphisme et le locus codant pour cet enzyme n'est donc pas utilisable (pas d'information disponible). Tous les individus produisent une enzyme qui a la même charge. On estime qu'un tiers seulement des mutations de l'ADN correspondant au gène d'un enzyme va donner une différence de charge suffisante pour être perçue par électrophorèse (Shaw, 1970).

Les taches révélées ne sont pas retrouvées au même endroit

Le locus correspondant à l'enzyme est polymorphe (plusieurs allèles). Plusieurs cas illustrés dans la figure 3 peuvent se présenter. Dans la figure 3, la situation décrite par le Locus I correspond au polymorphisme (plusieurs allèles) d'une enzyme monomérique, c'est-à-dire qu'une seule unité polypeptidique constitue l'enzyme fonctionnelle, celle décrite par le Locus II, représente un cas d'enzyme dimérique et celle du Locus II, une enzyme tétramérique.

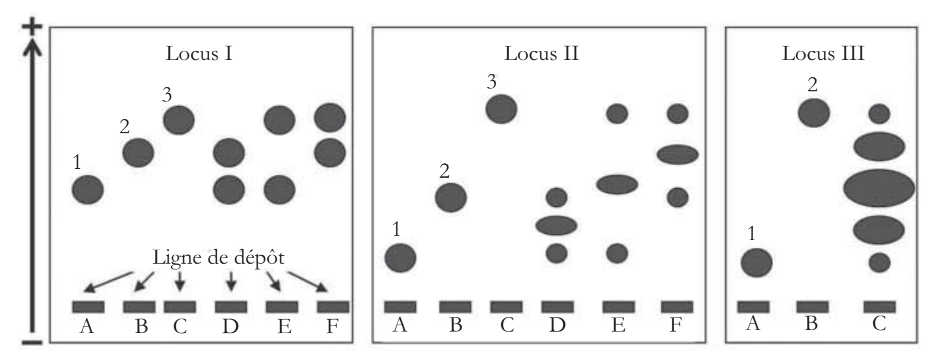


Figure 3
Représentation schématique des différents types de profils rencontrés avec des loci enzymatiques polymorphes. Le Locus I montre trois allèles différents (1, 2 et 3) et l'enzyme correspondante est monomérique puisque les hétérozygotes présentent deux bandes (ou taches). Le Locus II correspond à une enzyme dimérique avec trois allèles également. Dans ce cas, les hétérozygotes présentent trois bandes (ou taches), une tache pour chacun des deux homodimères et une tache centrale et plus importante correspondant à la combinaison des deux ou hétérodimère. Le Locus III correspond à une enzyme tétramérique avec deux allèles. Les taches des hétérodimères sont toujours plus grosses que celles des homodimères, car statistiquement plus probables (il est facile de le vérifier en construisant un tableau). L'interprétation génotypique de ces différents loci devrait donc être 1/1, 2/2, 3/3, 1/2, 1/3 et 2/3 pour A, B, C, D, E et F aux loci I et II ; et 1/1, 2/2 et1/2 pour A, B et C au locus III.

Autres cas

Une même fonction enzymatique peut être assurée par plusieurs loci (gènes). Dans le cas de deux loci, il y aura donc deux types de bandes à interpréter. La figure 4 donne un exemple d'une enzyme correspondant à deux loci avec un locus monomorphe et l'autre, monomérique et polymorphe à deux allèles. Cependant, il existe des cas où les deux loci sont monomorphes ou polymorphes tous les deux.

Commentaires sur les allozymes

Les lecteurs soucieux d'approfondir leurs connaissances sur les techniques liées à l'électrophorèse des protéines trouveront beaucoup plus d'informations dans Pasteur *et al.* (1987) et Ben Abderrazak *et al.* (1993).

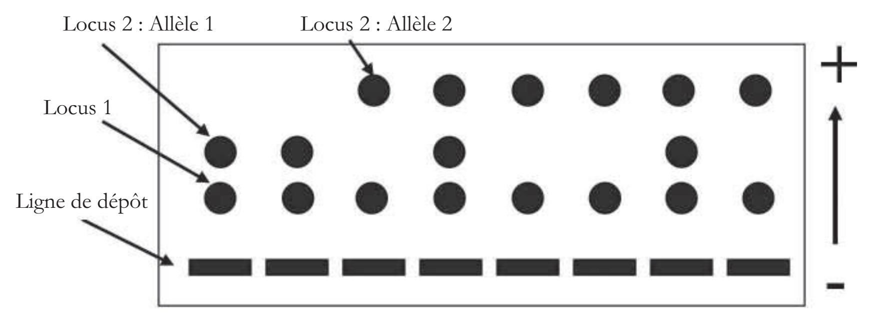


Figure 4
Cas d'une enzyme monomérique codée par deux loci différents, dont un (celui ayant le moins migré) est monomorphe et l'autre polymorphe avec deux allèles.

Les allozymes représentent ce qu'il y a de plus économique en temps et en argent. Malheureusement, ils sont rarement polymorphes, notamment chez les organismes parasites, et requièrent de travailler sur du matériel frais (maintien de la chaîne du froid), ce qui s'avère souvent difficile, en particulier dans les pays du Sud. Le matériel biologique à utiliser doit se trouver en quantité suffisante par individu, ce qui est souvent difficile avec les organismes parasites, souvent de taille modeste, si ces derniers ne sont pas cultivables (en les clonant). Ces loci correspondent à des séquences codantes de l'ADN. Leur polymorphisme est, de par ce fait, souvent suspecté de ne pas être entièrement neutre (Jarne et Lagoda, 1996). Or la neutralité (absence de sélection) est une hypothèse qui s'avérera importante (voir plus bas). Ces différents arguments permettent d'expliquer pourquoi les allozymes sont aujourd'hui peu utilisés en génétique des populations naturelles et en particulier, en épidémiologie moléculaire avec cependant quelques exceptions notables telles que celles représentées par de récentes études sur différents types d'organismes tels que des cafards (Corley *et al.*, 2001), des diptères (Niklasson *et al.*, 2004), des champignons pathogènes (Arnaviehle *et al.*, 2000 ; Badoc *et al.*, 2002 ; De Meeûs *et al.*, 2002b ; Nébavi *et al.*, 2006), et des parasites kinétoplastidés et leurs vecteurs (Barnabé *et al.*, 2000 ; Borges *et al.*, 2000 ; Hide *et al.*, 2001 ; Brenière *et al.*, 2003 ; Njiokou *et al.*, 2004).

Les microsatellites

Les microsatellites correspondent à des courtes séquences d'ADN répétées en tandem. Le plus généralement, sont considérés comme microsatellites les motifs répétés suivants :

– dinucléotides : exemple …GTGTGTGTGTGT…

– trinucléotides : exemple …CATCATCATCATCAT…

– tétranucléotides : exemple …GATAGATAGATAGATAGATAGATAGATA…

Les mononucléotides sont rarement utilisés, car trop instables et les pentanucléotides (et au-delà) deviennent plus rares. Au-delà, on a à faire à ce qui est appelé des minisatellites. La structure particulière de ces séquences les rend très susceptibles à la mutation. C'est-à-dire que les taux de mutation des séquences microsatellites seront souvent très élevés et, en conséquence, leur polymorphisme en populations naturelles sera lui aussi élevé en général (10^{-3}, 10^{-4}) (Ellegren, 2000 ; Balloux et Lugon-Moulin, 2002 ; Ellegren, 2004). Ce polymorphisme correspond donc à une variation dans le nombre de copies du motif de base. Par exemple avec $(AC)_n$, où n représente le nombre de répétitions, si on a n = 5, 6 ou 10, on a 3 allèles. Par ailleurs, ce sont souvent des séquences non codantes, sauf peut-être les trinucléotides qui correspondent potentiellement à des codons répétés. Les microsatellites impliqués dans des maladies génétiques (X fragile, dystrophie myotonique, maladie de Huntington…) sont d'ailleurs toujours des trinucléotides (Ashley et Warren, 1995 ; Fernández-López *et al.*, 2004). Or le polymorphisme de séquences non codantes a toutes les chances d'être neutre, sauf si le microsatellite en question se trouve par malchance à proximité d'un gène, ou dans un gène (intron) ayant subi un événement récent de sélection. Un autre avantage des microsatellites est qu'ils correspondent à des séquences relativement courtes d'ADN. En tant que tels, ils peuvent être amplifiés par PCR à partir de tissus conservés dans l'alcool pendant une durée assez longue et dans n'importe quelle (mauvaise) condition (en principe). L'amplification par PCR nécessite la connaissance des deux séquences flanquantes du locus où sont choisies les deux séquences complémentaires des amorces (ou *primers* en anglais). Pour ce faire, soit quelqu'un d'autre a déjà défini ces séquences et mis au point les techniques de PCR pour l'espèce étudiée (ou éventuellement sur une espèce proche), soit vous avez vous-même défini ces séquences à partir d'une banque génomique séquencée préexistante, soit vous avez constitué vous-même une banque génomique suivie d'un *screening* approprié (recherche de séquences microsatellites à l'aide de sondes) sur le détail duquel je ne m'étendrai pas. Le lecteur pourra cependant se référer aux protocoles détaillés disponibles sur internet. Citons à titre d'exemple celui de Toonen (1997) qui semble assez complet. Admettons que nous ayons ces fameuses séquences amorces à notre disposition. L'extraction de l'ADN de chaque individu est suivie, à partir d'une partie (ou aliquote) de cet ADN, d'une amplification par PCR spécifique (grâce aux amorces) de la séquence voulue et du marquage (radioactif ou fluorescent) du produit d'amplification. Une électrophorèse sur un support donné (gel de polyacrylamide, solution tampon) va ensuite permettre de discriminer les différents allèles en fonction de leur taille. Il y aura donc autant de bandes ou de pics (séquenceur automatique) différents qu'il y a d'allèles et tous les hétérozygotes auront deux bandes ou deux pics (fig. 5). Notons que si la séquence flanquante mute au niveau d'un des deux *primers* de telle sorte que l'appariement ne se fait plus, alors l'allèle correspondant ne sera plus amplifié. On parle alors d'un allèle nul. Un allèle nul ne peut, par définition, pas être détecté sauf à l'état homozygote (absence de bande). On peut aussi dire qu'il est récessif vis-à-vis

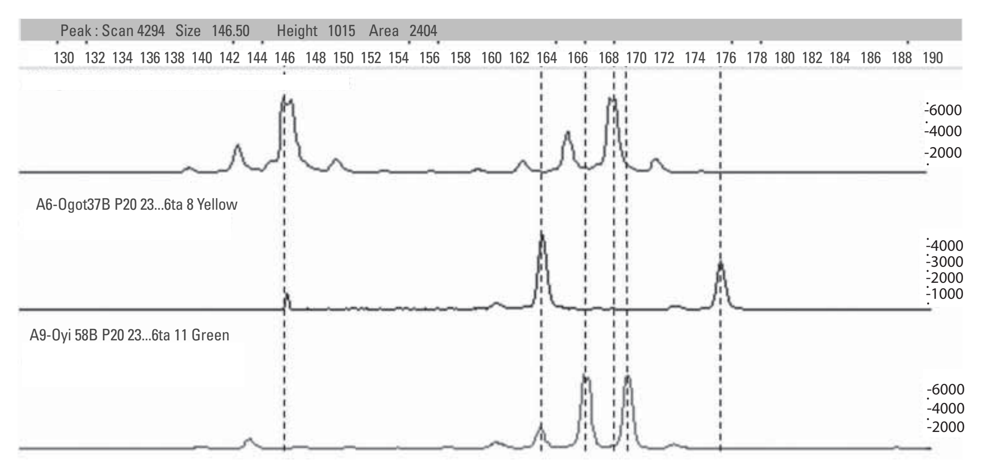


Figure 5
Exemple de profils obtenus pour des loci microsatellites dinucléotidiques sur séquenceur automatique. Les profils montrés correspondent à ceux obtenus à partir d'un oocyste de *Plasmodium falciparum* (agent de la forme la plus sévère de la malaria) et pour trois loci avec trois couleurs différentes, tous trois hétérozygotes. Le locus bleu présente un hétérozygote avec deux allèles 46 et 168, le noir est 164/176 et le vert est 166/170. Le nom des allèles correspond ici directement à la taille du produit obtenu après PCR spécifique.

des allèles non nuls (qui sont donc dominants). Nous reviendrons plus tard sur les allèles nuls.

Les loci microsatellites sont considérés comme étant en général très polymorphes, codominants, abondants dans (presque) tous les génomes et relativement aisés à manipuler (Lehmann *et al.*, 1996). Grâce à l'utilisation de l'outil PCR et des derniers progrès faits en la matière, il est possible de travailler à partir de quantités infimes de matériel biologique, comme le montre le travail de Razakandrainibe *et al.* (2005) où chaque oocyste de *Plasmodium falciparum* est analysé au niveau de sept marqueurs microsatellites. Ces arguments font des marqueurs microsatellites des outils de choix, sinon les meilleurs, pour les études de génétique de populations et en particulier, des populations de pathogènes (épidémiologie moléculaire). Le lecteur désireux de s'informer davantage sur les propriétés des microsatellites et leurs diverses applications est invité à consulter l'ouvrage édité par Goldstein et Schlötterer (1999).

2 Concepts de base en génétique des populations

CALCUL DES FRÉQUENCES ALLÉLIQUES À PARTIR D'UN ÉCHANTILLON

À partir de maintenant, nous considérerons, sauf si précisé, que nous travaillons sur un locus codominant (on distingue les hétérozygotes des homozygotes) avec deux allèles chez un organisme diploïde. Admettons que nous ayons génotypé N individus de cet organisme dans un site que nous supposons contenir une population. Parmi ces N individus, N_{11} se sont révélés être de génotype 1/1, N_{12} 1/2 et N_{22} 2/2. Notez que N est nécessairement égal à $N_{11} + N_{12} + N_{22}$. Soit p_1 et p_2, les fréquences des allèles 1 et 2 respectivement dans l'échantillon de N ndividus. Il y a $2N$ allèles en tout puisque nous sommes chez des diploïdes. Il y a $2N_{11}$ et N_{12} allèles 1 chez les 1/1 et 1/2 respectivement et $2N_{22}$ et N_{12} allèles 2 chez les 2/2 et 1/2 respectivement. La fréquence des allèles 1 et 2 dans l'échantillon est donc :

$$p_1 = \frac{2N_{11} + N_{12}}{2N} = \frac{N_{11} + \frac{1}{2}N_{12}}{N} \qquad (1)$$

et

$$p_2 = \frac{2N_{22} + N_{12}}{2N} = \frac{N_{22} + \frac{1}{2}N_{12}}{N} \qquad (2)$$

Notez que ces valeurs sont aussi des estimations des fréquences alléliques de la population échantillonnée et que, grâce à la codominance du marqueur, nous n'avons pas eu à faire d'hypothèse pour estimer ces fréquences alléliques (en fait si, laquelle ? Lire la réponse 1 en fin de volume).

CONFORMITÉ AVEC LES PROPORTIONS D'HARDY-WEINBERG

Les hypothèses d'Hardy-Weinberg

Hardy, mathématicien britannique, et Weinberg, médecin allemand, ont émis le même modèle de façon indépendante (Hardy, 1908 ; Weinberg, 1908). Ce modèle,

communément appelé « Équilibre d'Hardy-Weinberg », sert de base à une immense partie des études de génétique des populations.

Les hypothèses de ce modèle sont les suivantes :

– taille de population infinie ;

– pas de mutation ;

– pas de migration ;

– pas de sélection (neutralité) ;

– reproduction sexuée avec rencontre au hasard des gamètes (panmixie) ;

– pas de chevauchement de générations.

L'équilibre d'Hardy-Weinberg

Soit un locus à deux allèles 1 et 2 de fréquences p_1 et p_2 dans une telle population. Alors, puisque les gamètes se rencontrent au hasard, nous pouvons construire le tableau 1 qui décrit la rencontre des gamètes de la façon suivante :

Tableau I
Tableau de rencontre au hasard des gamètes. Les génotypes formés sont entre parenthèses.

			Gamètes de type	
			1	2
		Fréquences	p_1	p_2
Gamètes de type	1	p_1	p_1^2 (1/1)	$p_1\,p_2$ (1/2)
	2	p_2	$p_1\,p_2$ (2/1)	p_2^2 (22)

Nous attendons donc, dans les zygotes ainsi formés, les proportions de génotypes :

p_1^2, 2 $p_1\,p_2$ et p_2^2 pour 1/1, 1/2 et 2/2 respectivement. Et puisqu'il n'y a ni sélection, ni mutation, ni migration et que la population est infinie (pas de dérive aléatoire des fréquences alléliques), ces proportions resteront les mêmes chez les adultes de la génération suivante. En se rappelant que la somme $p_1 + p_2$ est nécessairement égale à 1, nous démontrons facilement que les nouvelles fréquences alléliques seront donc, en appliquant l'équation (1) :

$$p'_1 = \frac{p_1^2 + \frac{1}{2}2p_1p_2}{p_1^2 + 2p_1p_2 + p_2^2} = \frac{p_1^2 + p_1p_2}{(p_1 + p_2)^2} = \frac{p_1(p_1 + p_2)}{(1)^2} = \frac{p_1(1)}{1} = p_1$$

et donc

$$p'_2 = p_2$$

C'est ce que l'on appelle l'équilibre (car rien ne bouge) de Hardy-Weinberg.

Que se passe-t-il si nous relâchons chacune des hypothèses du modèle de Hardy-Weinberg l'une après l'autre ?

RELAXATION DES HYPOTHÈSES DE HARDY-WEINBERG

La population est de taille finie

Prenons un exemple extrême d'une population de taille 2. Admettons qu'à la génération 0, les deux individus sont hétérozygotes 1/2. La fréquence des deux allèles est donc ½. Ils fabriquent des gamètes qui se rencontrent au hasard pour former ¼, ½ et ¼ de zygotes 1/1, 1/2 et 2/2 respectivement (nous considérons ici un cas où le nombre de zygotes produit est très grand). Il faut reconstituer à partir de ces zygotes une population d'adultes de taille 2 (phénomène démographique appelé régulation). J'ai $(¼)^2$ de choisir au hasard deux 1/1, 2(¼)(½) d'avoir un 1/1 et un 1/2, $(½)^2$ d'avoir deux 1/2, $2(¼)^2$ de choisir un 1/1 et un 2/2, 2(½)(¼) d'avoir un 1/2 et un 2/2 et $(¼)^2$ d'avoir deux 2/2. Ce qui veut dire qu'à la génération suivante, j'ai 5/8 chances d'obtenir une configuration avec des fréquences alléliques différentes de celles de la génération 0, et dans ⅛ cas de fixer définitivement la population en 11 ou en 22. C'est ce que l'on appelle la dérive génétique. Dans une population de taille finie, le hasard modifie les fréquences alléliques d'une génération à l'autre. Ceci veut dire que s'il n'y avait rien d'autre (mutation, migration, sélection), aucun polymorphisme, à aucun locus, ne pourrait se maintenir dans les populations naturelles. Notons que le phénomène est d'autant plus rapide que les populations sont de petites tailles.

Il y a mutation

Cela correspond aux erreurs de copies lors de la duplication de l'ADN durant la construction des gamètes. Il existe plusieurs types de mutation.

Mutation récurrente

Une mutation récurrente correspond à la transformation d'un allèle donné en un autre allèle donné (par exemple, 1- > 2). C'est le cas de la plupart des mutations délétères comme l'albinisme par exemple, qui intervient avec la fréquence (taux de mutation) de 2.5×10^{-5} environ (Hedrick, 2003), et ces mutations correspondent en général à une perte de fonction. Le taux de mutation en est en général assez bas (10^{-5}, 10^{-6}) et la mutation reverse est en général très faible et négligeable (car il faut réparer exactement ce qui a été perdu), de l'ordre de 10^{-8}.

Modèle de mutation en nombre fini d'allèles ou KAM (K Alleles Model)

La mutation transforme l'allèle d'origine vers n'importe quel type d'allèle parmi les *K* possibles. Si par exemple, on regarde le polymorphisme au niveau d'une seule paire de base, on aura $K = 4$ (A, T, G ou C) (à vous de trouver pourquoi cet exemple n'est pas très bon, sinon reportez-vous à la réponse 2 en fin de volume). Pour les allozymes, on a en général entre 1 et 10 allèles visibles. Pour d'autres marqueurs, *K* peut être très grand. À cause de ce nombre limité d'allèles possibles, il peut arriver que deux allèles soient identiques sans pour autant partager une origine ancestrale commune proche dans le temps (*coancestry* en anglais). On dit de ces allèles qu'ils sont identiques par état et non par descendance et on appelle ce phénomène homoplasie.

IAM ou Infinite Allele Model

La mutation transforme l'allèle d'origine vers un allèle nouveau (jusqu'alors inexistant) et indépendant de l'état du précédent. Ce modèle est très utilisé en génétique des populations théoriques, car il est plus simple à manipuler sans modifier considérablement les résultats par rapport au KAM (si *K* suffisamment grand). Dans ce modèle, il n'y a pas d'homoplasie et deux allèles identiques par état le sont également par descendance.

SMM ou Stepwise Mutation Model

Le SMM (Kimura et Ohta, 1978) est un mode de mutation invoqué pour les marqueurs microsatellites. La mutation correspond ici à l'ajout ou au retrait d'une répétition par rapport à l'allèle d'origine. Il est évident que ce type de mutation va générer beaucoup d'homoplasie. Il en résulte également qu'une ressemblance de taille traduira également une proximité ancestrale probable. Il existe aussi des modèles panachés entre KAM et SMM, comme le TPM (*Two Phases Model*) avec une certaine proportion de SMM et le complément en KAM avec une variance de taille d'allèles donnée.

Conclusion sur la mutation

Quoi qu'il en soit, la mutation va bien évidemment modifier les fréquences alléliques des loci qu'elle affecte. Il faut noter cependant que les taux de mutation sont en général assez bas (sauf pour quelques microsatellites) et que la mutation seule ne peut donc pas représenter une force très puissante de l'évolution des populations. Il n'y aurait néanmoins pas d'évolution sans mutation, car c'est la seule source de nouveauté héritable, et, associée à la dérive et à la sélection, la mutation représente en effet la clé de l'évolution et de la structure génétique des populations.

Migration

Les populations naturelles ne sont pas isolées totalement les unes des autres. Elles reçoivent donc des propagules venant d'autres populations plus ou moins éloignées

et présentant, à l'ensemble du génome, des fréquences d'allèles plus ou moins différentes. Ces propagules peuvent être des individus adultes, larvaires, des gamètes (pollen) ou des spores. Ils peuvent donc être haploïdes ou diploïdes. La migration peut être forte. Elle a tendance à homogénéiser les populations entre elles (génétiquement). C'est donc une force potentiellement majeure de l'évolution des populations. Notons ici que, comme nous le verrons plus loin, associée à la dérive et à la mutation, la migration peut conduire, en population structurée, à l'établissement d'un polymorphisme stable (équilibre) d'une génération à l'autre et sans l'intervention d'une quelconque forme de sélection. On peut même observer, dans certains types de populations structurées, l'établissement d'un cline géographique des fréquences alléliques.

Sélection

La sélection est évidemment une force majeure de l'évolution. Elle peut prendre de multiples formes et peut affecter un, quelques-uns ou plusieurs loci en même temps et agir à différents niveaux (génomique, individuel, populationnel…) avec différents effets et interactions. Il s'agit donc d'un domaine d'investigation très large. Nous ne passerons en revue que quelques exemples parmi les plus simples et les plus utiles à la suite de notre propos.

Sélection directionnelle

Comme son nom l'indique, la sélection directionnelle tend à augmenter ou diminuer la fréquence d'un allèle dans la population, en affectant la survie ou la reproduction des porteurs de cet allèle pour le locus concerné. La vitesse du processus dépend de la force de la sélection, de la dominance (ou récessivité) de l'allèle vis-à-vis de la sélection, du système de reproduction et de la taille de la population. Sans mutation, l'aboutissement de cette sélection est la fixation de l'allèle le plus favorable à la survie et/ou reproduction des individus qui le portent. Cette sélection n'est détectable qu'expérimentalement ou par des études corrélatives car, seule, elle n'affecte pas ou très peu le schéma génotypique p_1^2, $2\ p_1\ p_2$ et p_2^2 d'Hardy-Weinberg. Seules les fréquences alléliques changent. Cette sélection peut cependant modifier le degré de différenciation entre populations différentes aux loci concernés. En fonction des cas, elle peut diminuer la différenciation (sélection convergente) quand la direction de la sélection est la même d'un site à l'autre. Elle peut au contraire augmenter cette différenciation lorsque la direction de sélection est variable d'un site à l'autre (sélection divergente ou disruptive) (voir par exemple De Meeûs *et al.*, 1993 ; De Meeûs et Goudet, 2000 ; De Meeûs, 2000). Normalement, cette forme de sélection n'est perceptible qu'aux loci (gènes) concernés et à ceux qui leur sont liés (auto-stop) et pas aux autres marqueurs. C'est donc un phénomène locus spécifique.

Sous-dominance

C'est le nom qu'on lui donne même si ce n'est guère explicite pour ne pas dire très mal choisi. Il s'agit d'une sélection qui défavorise les hétérozygotes. Cette forme de sélection conduit théoriquement à l'élimination de l'allèle le moins fréquent. En effet, s'il y a panmixie, l'allèle le plus rare sera le plus souvent hétérozygote (vous n'avez qu'à vérifier cela dans les proportions attendues chez les zygotes) et donc le plus souvent défavorisé. Il existe peu ou pas d'exemples de sous-dominance. L'exemple le plus connu qui s'en rapproche le plus est le cas du système Rhésus (Hartl et Clark, 1989). Dans ce système, les Rh+Rh- sont en moyenne défavorisés par rapport aux Rh+Rh+ ou Rh-Rh-, car les femmes Rh-Rh- ont plus de chance de perdre un enfant (quand ce dernier est Rh+Rh-). Le maintien d'un tel polymorphisme dans les populations humaines est assez surprenant. Tant que le polymorphisme persiste, la signature d'un tel processus est un déficit en hétérozygotes, par rapport aux attendus de Hardy-Weinberg, chez les adultes, et donc un excès d'homozygotes, pour le locus concerné (et seulement lui). Avec deux allèles (1 et 2) de fréquences respectives p_1 et p_2, cela donne les fréquences génotypiques : $p_1^2 + p_1p_2F_{IS}$, $2p_1p_2(1 - F_{IS})$ et $p_2^2 + p_1p_2F_{IS}$, pour 1/1, 1/2 et 2/2 respectivement, avec F_{IS} le déficit en hétérozygotes (voir plus loin).

Super-dominance

Là non plus, le terme n'est pas très heureux, mais c'est ainsi. Ici, ce sont les homozygotes qui sont moins favorisés (ou avantage de l'hétérozygote). Dans ce cas, la population tend à converger vers un équilibre stable des fréquences alléliques au locus concerné (et seulement lui). Il existe encore une fois peu d'exemples naturels de ce phénomène. Les plus connus sont la résistance à la malaria des patients hétérozygotes pour la drépanocytose (ou anémie falciforme) (Ridley, 1996). Il y a deux allèles au locus responsable. Le premier allèle (+) dit sauvage, et le second (-) dit mutant. Les individus -/- sont atteints d'une maladie génétique grave (survie et reproduction très compromises), les individus +/+ sont normaux, mais les individus +/- sont en moins bonne santé que les +/+ sauf dans les populations soumises à une forte pression par *Plasmodium falciparum* (l'agent le plus virulent de la malaria). Dans ce dernier cas, les +/+ ont des taux de survie inférieurs à celui des +/-, qui eux-mêmes survivent mieux que les -/- (qui sont très malades, quelles que soient les conditions), il y a super-dominance. Notons que ces modes de résistance sont coûteux en termes de zygotes produits, puisqu'une grande partie des individus produits à chaque génération sont homozygotes et donc moins bien adaptés. Une échappatoire à ce travers peut provenir du système de reproduction s'il fait en sorte qu'une majorité d'hétérozygotes soient issus de la reproduction. Ceci se traduirait par un coût au niveau reproductif (choix du conjoint) et les individus hétérozygotes produits sont tous condamnés à une descendance imparfaite. La signature de ce phénomène sur des marqueurs génétiques est bien évidemment la présence d'excès d'hétérozygotes par rapport aux attendus de

Hardy-Weinberg, pour le seul locus concerné par cette sélection, bien évidemment, et éventuellement pour les loci les plus liés au gène sous sélection (auto-stop).

La sélection fréquence-dépendante

On l'appelle aussi sélection apostatique (avantage du rare ou apostat) : plus un allèle est rare et plus l'individu qui le porte a de chances de survivre et/ou de se reproduire. Les exemples sont multiples. Les plus connus concernent ce qui a trait aux systèmes immunitaires et à la sélection sexuelle (Schierup *et al.*, 2001). Chez le trèfle, par exemple, on connaît un locus d'auto-incompatibilité possédant une multitude d'allèles différents (Lawrence, 2000). Une fleur de trèfle ne peut être fécondée que par un pollen ne possédant aucun des deux allèles présents chez la fleur à ce locus. Il en résulte que les plantes sont nécessairement toutes hétérozygotes à ce locus et que tout mutant ou migrant possédant un allèle nouveau sera fortement favorisé (il peut féconder, et être fécondé par, tout le monde). Le système MHC (Complexe majeur d'histocompatibilité) des mammifères ou HLA (Antigène lymphocytaire humain) chez l'homme, fonctionne selon un principe équivalent puisqu'un couple dont le HLA est trop similaire est stérile, et qu'il y a manifestement des attirances dépendantes de la différence entre le MHC des deux partenaires (Wedekind et Penn, 2000). Ici, la signature du phénomène est facile à repérer, puisque les loci impliqués doivent avoir une hétérozygotie fixée ou au moins très élevée. D'autres exemples peuvent concerner des systèmes de résistance hôte/virulence parasite. C'est le cas des modèles de gène-pour-gène (avec coûts sélectifs) où seuls les parasites « virulents » peuvent infecter les hôtes « résistants », alors que les hôtes susceptibles peuvent aussi être envahis par les parasites « avirulents » ; c'est le cas aussi des modèles appelés « matching alleles » où chaque allèle de résistance de l'hôte ne permet l'invasion que d'un type de parasite porteur d'un allèle de virulence précis (se référer à Agrawal et Lively, 2002 pour une description plus détaillée de ces deux modèles). On conçoit que si on a par exemple deux types de parasites P1 et P2 et deux types d'hôtes H1 et H2, si seul H1 est compatible pour P1 et H2 pour P2, mais que ce parasite est létal pour l'hôte dans lequel il parvient à s'installer, on comprend bien que ce système fonctionnera de façon fréquence-dépendante. Ici, la signature de ce système au niveau du locus en tant que marqueur génétique ne sera pas évidente à mettre en évidence autrement que par des expériences ou des suivis dans le temps de tous les acteurs du système. La fréquence-dépendance aura souvent tendance à homogénéiser les fréquences alléliques des loci concernés sur une grande part de l'aire de répartition de l'espèce. Cependant, l'interaction avec les schémas de migration peut potentiellement complexifier ce schéma (Gandon *et al.*, 1996 ; Gandon, 2002 ; Morgan *et al.*, 2005).

Hétérosis

L'hétérosis (ou vigueur hybride) est un phénomène global qui affecte la totalité du génome. Il peut provenir d'une superdominance globale répartie sur de très nombreux loci du génome ou bien il résulte de la présence de nombreux allèles

délétères récessifs dans la population qui fait que plus un individu est hétérozygote au plus grand nombre de loci et plus sa valeur sélective croît (voir PRUGNOLLE *et al.*, 2004a). Ici, la signature génétique de ce phénomène correspond à un excès d'hétérozygotes sur l'ensemble des loci testés. Il convient cependant de pouvoir écarter les hypothèses alternatives, que nous aborderons plus loin, pouvant expliquer un excès d'hétérozygotie multilocus tels que la clonalité (BALLOUX *et al.*, 2003), l'existence de petites populations dioïques ou autoincompatibles (BALLOUX, 2004) avec ou sans biais de dispersion sexe-spécifique (PROUT, 1981 ; PRUGNOLLE et DE MEEÛS, 2002) ou les membres d'une même fratrie (individus issus de la même ponte) (CHEVILLON *et al.*, 2007a). Ce phénomène aura tendance à homogénéiser les fréquences alléliques entre différents sites (sous-populations) à tous les loci impliqués et donc potentiellement sur l'ensemble des loci du génome (auto-stop).

La sélection gamétique

La sélection gamétique donne un avantage à certains gamètes (spermatozoïdes plus performants). C'est une forme de sélection souvent négligée mais très puissante, comme en atteste le maintien de mutations délétères (même sub-létales) à des fréquences anormalement élevées (NUNNEY et BAKER, 1993).

Le régime de reproduction n'est pas panmictique

Ici, aussi plusieurs cas sont possibles.

Autofécondation

Ceci n'est bien sûr possible que chez des organismes hermaphrodites (*Taenia*, *Echinococcus*, *Fasciola*, *Plasmodium*) (nous ne parlerons pas ici de certains cas de parthénogénèse automictique). Imaginons que chez de tels organismes, une proportion s de gamètes est investie dans l'autofécondation et donc $1\text{-}s$ dans des croisements panmictiques. En reprenant notre locus à deux allèles de tout à l'heure, nous pouvons poser que D_n, H_n et R_n sont les fréquences des génotypes 1/1, 1/2 et 2/2 à la génération n respectivement, avec $D_n = N_{11}/N$, $H_n = N_{12}/N$ et $R_n = N_{22}/N$. Nous supposerons ici que N (taille de la population) est très grand. Ces individus se reproduisent. Quelles seront les fréquences génotypiques à la génération suivante ?

– Pour D_{n+1} : par autofécondation (proportion s des zygotes produits), seuls les 1/1, en proportion D_n, et les 1/2, en proportion H_n, de la génération n peuvent produire des 1/1. Dans ce cas, les 1/1 qui s'autofécondent ne produisent que des 1/1 (on suppose qu'il n'y a pas de mutation) et les 1/2 ne produisent par autofécondation que ¼ de 1/1 (le reste étant ½ de 1/2 et ¼ de 2/2). Par panmixie ($1\text{-}s$ des zygotes), on a vu que la proportion de 1/1 produite est de p_1^2 (la fréquence de l'allèle 1 chez les zygotes n'a pas de raison d'être différente de celle de la population). On a donc :

$$D_{n+1} = s\,[D_n + \tfrac{1}{4}\,H_n] + (1 - s)\,p_1^2$$

– Pour H_{n+1} : seuls les hétérozygotes (H_n) peuvent produire d'autres hétérozygotes par autofécondation (s) (pour moitié, car le reste se répartit en ¼ de 1/1 et ¼ de 2/2, comme on l'a vu), et la panmixie (1 - s) en produit $2p_1 p_2$, donc :

$$H_{n+1} = s\,[\tfrac{1}{2} H_n] + (1 - s)\, 2p_1 p_2$$

– Pour R_{n+1} : on a la situation symétrique à celle de D_{n+1}, à savoir :

$$R_{n+1} = s\,[R_n + \tfrac{1}{4} H_n] + (1 - s)\, p_2^2$$

Nous avons maintenant toutes les informations nécessaires pour calculer la fréquence d'équilibre des hétérozygotes, si elle existe. À l'équilibre, plus rien ne bouge (par définition), et nous obtenons donc $H_{n+1} = H_n = H_e$. Nous pouvons alors poser :

$H_e = s\,[\tfrac{1}{2} H_e] + (1 - s)\, 2p_1 p_2$ et donc

$H_e - s\,[\tfrac{1}{2} H_e] = (1 - s)\, 2p_1 p_2$, d'où

$H_e\,[1 - \tfrac{1}{2} s] = (1 - s)\, 2p_1 p_2$, d'où

$$H_e = \frac{(1-s)2p_1p_2}{1-\frac{1}{2}s} = \frac{\left(1-\frac{1}{2}s-\frac{1}{2}s\right)2p_1p_2}{1-\frac{1}{2}s} = 2p_1p_2\left(1-\frac{\frac{1}{2}s}{1-\frac{1}{2}s}\right)$$

$$H_e = 2p_1p_2\left(1-\frac{s}{2-s}\right) \qquad (3)$$

et donc pour D_e et R_e on a de la même façon :

$$D_e = p_1^2 + p_1p_2\frac{s}{2-s} \qquad (4)$$

et

$$R_e = p_2^2 + p_1p_2\frac{s}{2-s} \qquad (5)$$

D'après l'équation (3), on voit que si $s = 0$ on retrouve Hardy-Weinberg. Si $s = 1$, on obtient $H_e = 0$, ce qui revient à dire qu'il ne reste pas d'hétérozygotes à l'équilibre, seulement p_1 1/1 et p_2 2/2 (facile à vérifier avec les équations 4 et 5, sinon allez voir la réponse 3). C'est ce qui se passe par exemple chez *Taenia solium* (Kunz, 2002 ; De Meeûs *et al.*, 2003). Si s est entre 0 et 1, il y aura un déficit plus ou moins important d'hétérozygotes. Il est très important de noter que la même signature de l'autofécondation est attendue à tous les loci étudiés (signature génomique).

Le fait qu'un organisme soit hermaphrodite et puisse s'autoféconder n'implique pas nécessairement que ses populations ne soient pas panmictiques. Par exemple, en utilisant des marqueurs microsatellites, Hurtrez-Boussés *et al.* (2004) ont trouvé que les populations de la grande douve du foie *Fasciola hepatica*, plathelminthe hermaphrodite, montraient des fréquences génotypiques conformes à l'attendu sous panmixie. En panmixie, on attend en effet que $1/N$ des zygotes produits le soient par autofécondation (Rousset, 1996). Ce sont plutôt les organismes à sexes séparés qui ne sont jamais panmictiques entièrement (les gènes contenus dans les femelles ne peuvent s'associer qu'à ceux contenus dans les mâles). Ceci n'a vraiment

d'importance que dans les petites populations. Chez les espèces dioïques ou chez les hermaphrodites autoincompatibles, on s'attend à détecter des excès d'hétérozygotes par rapport à l'attendu sous les hypothèses de Hardy-Weinberg (Balloux, 2004). Des excès d'hétérozygotes plus ou moins prononcés sont donc attendus chez de nombreuses espèces parasites tels que les schistosomes (dioecie) ou les monogènes (monoïques largement autoincompatibles), ce qui a en effet été documenté pour *Schistosoma mansoni* (Prugnolle *et al.*, 2002).

Les croisements systématiques entre apparentés

Chez la guêpe parasitoïde *Nasonia vitripennis*, la femelle pond plusieurs œufs (frères-sœurs) dans une même chenille. Ceci a tendance à favoriser les croisements entre frères et sœurs (Shuker *et al.*, 2004). Dans certaines populations, c'est même la règle. Ce type de reproduction existe ou a existé de façon marginale dans l'espèce humaine pour certains membres de familles royales ou impériales (pharaons, rois européens). Le résultat est identique au précédent même si moins efficace (voir la figure 6). On obtient des déficits en hétérozygotes à tous les loci par rapport aux attendus sous l'hypothèse de panmixie.

L'homogamie

Ici, les individus de même génotype préfèrent s'accoupler entre eux ou la compatibilité entre gamètes est augmentée par la ressemblance génétique. Les conséquences sont identiques à l'autofécondation sauf qu'elles ne concernent que les gènes responsables du caractère (homogamie), et ceux qui leur sont liés (auto-stop), qui voient la fréquence des hétérozygotes diminuer. S'il y a dominance pour le caractère (chaque génotype se reconnaît), la vitesse de perte d'hétérozygotie sera la même que pour l'autofécondation, alors que s'il y a dominance pour le caractère (les hétérozygotes et homozygotes dominants s'accouplent de leur côté et les homozygotes récessifs du leur), cette vitesse dépend des fréquences alléliques. Des caractères tels que la taille à la maturité sexuelle ou la résistance aux pathogènes ont presque toujours, au moins en partie, un déterminisme génétique. Or, il est prouvé que dans de nombreuses espèces, ces caractères conditionnent l'appariement assorti (*assortative mating*) des partenaires sexuels (Thomas *et al.*, 1995).

La figure 6 illustre une comparaison de l'efficacité, en termes de perte d'hétérozygotie, des différents régimes consanguins de reproduction décrits plus haut. Remarquons que l'autofécondation est la plus efficace, que les croisements frères/sœurs sont les moins rapides, mais rattrapent l'homogamie avec dominance sur la fin et que les plus lents sont les homogames dominants pour lesquels l'allèle dominant est le plus fréquent dans la population de départ.

L'hétérogamie

L'autoincompatibilité est une forme d'hétérogamie. Elle ne peut exister sans sélection fréquence-dépendante (voir p. 35). Notons qu'elle ne concerne que les loci

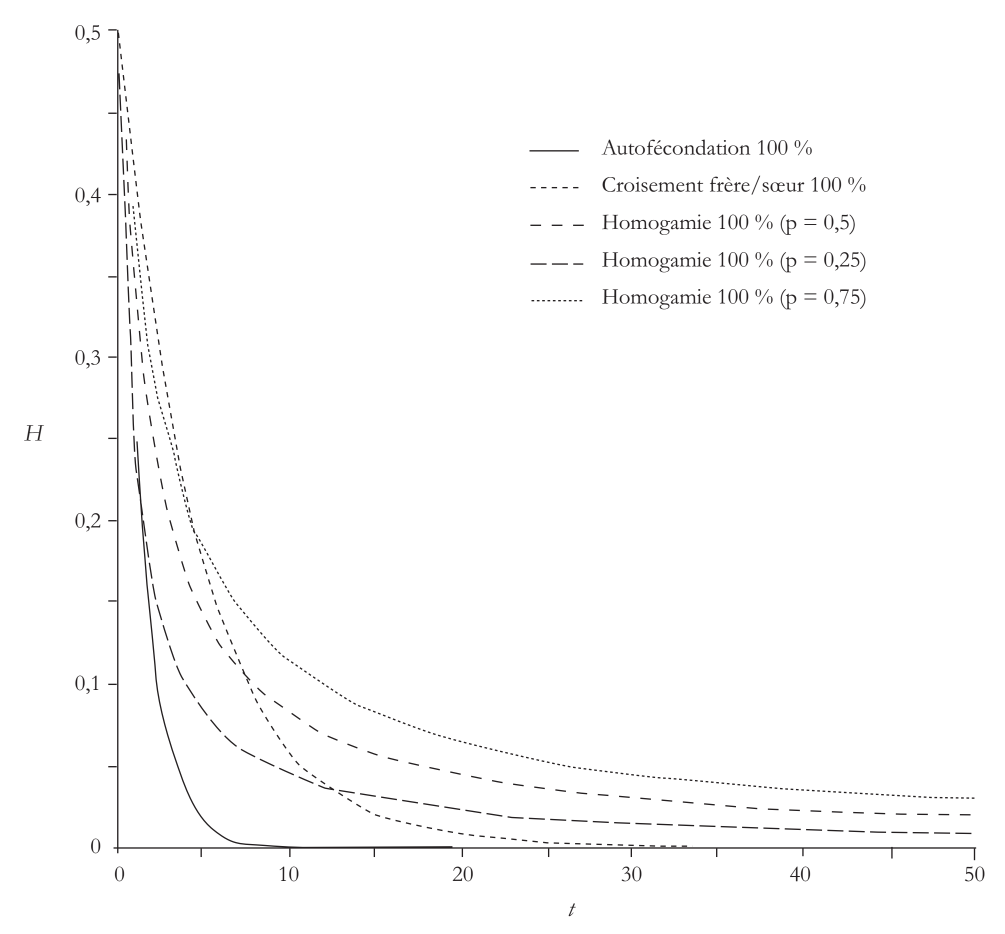


Figure 6
Perte d'hétérozygotie (*H*) en fonction du temps en générations non chevauchantes (*t*) pour différents types de croisements consanguins, pour deux allèles et pour une fréquence d'hétérozygote à t_0 de ½. Pour l'homogamie, les résultats sont donnés dans le cas où le premier allèle est dominant avec différentes fréquences (*p*) de cet allèle au locus concerné. Pour l'autofécondation et les croisements frères sœurs, les valeurs concernent l'ensemble des loci alors que pour l'homogamie, elles ne sont pertinentes que pour les loci concernés par le phénomène. Dans tous les cas, le phénomène concerne 100 % des gamètes (*s* = 1).

responsables du caractère. Cependant, nous pouvons aussi envisager une relation épistatique de l'ensemble du génome sur un locus d'évitement de l'apparentement. En effet, la consanguinité peut entraîner un fardeau important, il n'est donc pas déraisonnable de penser qu'il peut être avantageux de choisir les partenaires sexuels les moins apparentés pour former des zygotes. Une descendance plus hétérozygote et donc plus variable pourrait aussi apporter un avantage face à des agresseurs (parasites). Dans ce cas, on s'attend à un excès d'hétérozygotes sur tous les loci utilisés pour cette reconnaissance. Ceci peut aussi être accompli par un biais sexe-spécifique

de dispersion dans une population structurée (seuls les mâles dispersent, par exemple) (Prugnolle et De Meeûs, 2002 ; Prugnolle *et al.*, 2003). Dans ce cas, les accouplements se font entre individus plus divergents génétiquement que s'il y avait panmixie. Cela entraîne alors, comme déjà évoqué, de légers excès d'hétérozygotes à tous les loci (Prout, 1981 ; Prugnolle et De Meeûs, 2002). Un exemple récent sur les otaries à fourrure de l'île de Géorgie (hémisphère sud) a mis en évidence un choix délibéré des femelles pour s'accoupler avec des mâles non apparentés et plus hétérozygotes[2] (Hoffman *et al.*, 2007).

La clonalité

Par définition la clonalité, ou reproduction asexuée, ou encore parthénogenèse apomictique, reproduit à l'identique les individus qui la pratiquent. Elle ne peut donc rien changer à elle seule. Cependant, en populations finies subissant la dérive génétique, avec un taux de mutation constant, des excès d'hétérozygotes sont attendus par rapport aux fréquences génotypiques de Hardy-Weinberg à l'échelle de l'ensemble du génome et on s'attend même au bout d'un certain temps à une hétérozygotie totale, sauf pour les allèles homoplasiques (Balloux *et al.*, 2003 ; De Meeûs et Balloux, 2005 ; De Meeûs *et al.*, 2006 ; De Meeûs *et al.*, 2007b).

Les générations se chevauchent

Ce phénomène ne va pas créer une perturbation seul, mais combiné à la dérive, il va créer une hétérogénéité dans la population (effet Wahlund, voir plus loin) qui va se traduire par un déficit en hétérozygotes proportionnel à l'intensité de la dérive. Cela aura aussi tendance à minimiser certaines mesures de différenciation entre populations telles que le F_{ST} (voir plus loin).

LA NOTION DE DÉFICIT EN HÉTÉROZYGOTES, DÉFINITIONS

Comme nous venons de le voir, la structure génotypique d'une population, ${p_1}^2$, $2\,p_1\,p_2$ et ${p_2}^2$, attendue sous les hypothèses de Hardy-Weinberg, peut être altérée par certaines formes de sélection et par le système de reproduction. Il va donc y avoir d'autres fréquences génotypiques observées, D_o, H_o et R_o pour les génotypes 1/1, 1/2 et 2/2 respectivement (pour le cas à deux allèles). Si on ne s'intéresse qu'aux causes dues au système de reproduction (qui affectent donc tous les loci), on voit que ce

[2] Les individus les plus hétérozygotes sont probablement ceux qui présentent le plus grand choix d'allèles disponibles. Dans l'un et l'autre cas, les descendants peuvent espérer une plus grande hétérozygotie.

qui est perdu ou gagné par les hétérozygotes est normalement équitablement restitué aux homozygotes, comme suggéré en p. 37 équations 4 et 5 :

$D_o = p_1{}^2 + p_1p_2F_{IS}$

$H_o = 2p_1p_2(1 - F_{IS}) = 2p_1p_2 - 2p_1p_2F_{IS} = H_e - H_eF_{IS}$

$R_o = p_2{}^2 + p_1p_2F_{IS}$

d'où on peut tirer que :

$$F_{IS} = \frac{H_e - H_o}{H_e} = 1 - \frac{H_o}{H_e} \qquad (6)$$

où F_{IS} représente donc le ratio d'hétérozygotie en plus ou en moins observé par rapport à l'hétérozygotie attendue (H_e) sous les hypothèses de Hardy-Weinberg. Ce nouveau paramètre, défini par Wright (WRIGHT, 1965) est appelé indice de fixation (F) des individus dans les sous-populations (s) ou déficit en hétérozygotes. Il varie entre – 1 et + 1. Les valeurs négatives correspondant donc à un excès d'hétérozygotes, les valeurs positives à un déficit en hétérozygotes et une valeur nulle correspondant donc à Hardy-Weinberg. Il est intéressant de noter que – 1 ne peut être atteint que par une population où tous les individus sont hétérozygotes pour les mêmes deux allèles (par exemple, 1/2), alors que + 1 signifie seulement qu'il n'y a aucun hétérozygote, et donc tous les homozygotes que l'on veut. Il y a donc une contrainte sur les fréquences alléliques pour les F_{IS} négatifs : pour F_{IS} = - 1 on a donc nécessairement deux allèles avec $p_1 = p_2 = ½$.

On peut donc exprimer les fréquences génotypiques en fonction du F_{IS} :

$$\begin{cases} D_o = p_1^2 + p_1 p_2 F_{IS} \\ H_o = 2p_1 p_2 (1 - F_{IS}) \\ R_o = p_2^2 + p_1 p_2 F_{IS} \end{cases} \qquad (7)$$

ce qui correspond aux formules généralisées de Wright.

Nous pouvons donc calculer un déficit en hétérozygotes standardisé, indépendant des fréquences alléliques et donc comparable d'un locus à l'autre et d'une étude à l'autre. Prenons par exemple les effectifs génotypiques suivants : N_{11} = 15, N_{12} = 10 et N_{22} = 20, issus du génotypage allozymique d'une enzyme quelconque d'un échantillon de vers hermaphrodites prélevés dans un intestin de mammifère. En utilisant les équations (1) et (2), nous pouvons calculer les fréquences alléliques : p_1 = 0,44, $p_2 = (1 - p_1)$ = 0,56. D'où nous pouvons tirer, en utilisant (6) :

$$F_{IS} = 1 - \frac{H_o}{H_e} = \frac{\frac{N_{12}}{N}}{2p_1p_2} = 1 - \frac{\frac{10}{15+10+20}}{2 \times 0.44 \times 0.56} = 0{,}55$$

Ce résultat se traduit par le fait qu'il manque 55 % des hétérozygotes attendus sous l'hypothèse de panmixie. Si on fait l'hypothèse que ce déficit vient de l'autofécondation,

on peut utiliser les équations (3) et (7) pour estimer le taux d'autofécondation conduisant au F_{IS} observé. En effet, on voit bien qu'en combinant ces deux équations, on obtient :

$$F_{IS} = \frac{s}{2-s}$$

D'où on tire facilement que :

$$s = \frac{2F_{IS}}{1+F_{IS}} \qquad (8)$$

Nous avons ici un premier exemple d'inférence possible à l'aide de marqueurs moléculaires. La connaissance du déficit en hétérozygotes, en supposant que ce dernier ne vient que du régime de reproduction et qu'on est à l'équilibre génotypique, permet d'estimer la proportion d'autofécondation pratiquée par la population étudiée. Ceci a par exemple permis d'estimer ce taux d'autofécondation dans les populations de lymnées tronquées, escargot aquatique hôte intermédiaire de la grande douve du foie (*s* estimé entre 0,8 et 1) (Meunier *et al.*, 2004a). Si la population n'est pas à l'équilibre génotypique, il s'agit alors de valeurs minimales nécessaires pour expliquer les fréquences génotypiques observées. Dans le cas des lymnées tronquées, cela ne change pas grand-chose d'ailleurs, car on est proche du maximum possible.

Dans le cas de loci à plus de deux allèles, il va exister autant de F_{IS} que d'allèles. On comprend facilement que la multiplicité des F_{IS} ne va pas favoriser l'interprétation des processus qui conduisent aux fréquences génotypiques observées (comme le taux d'autofécondation). On peut calculer un F_{IS} moyen sur l'ensemble des allèles. On peut faire la moyenne non pondérée, mais la méthode la plus populaire, et la meilleure à mon sens, correspond à la moyenne des F_{IS} par allèle pondérée par le produit des fréquences alléliques $p_i(1 - p_i)$. Ce type de pondération permet de donner le maximum de poids aux allèles de fréquences intermédiaires, et peu de poids aux allèles rares.

Une mesure du F_{IS} sur un seul locus est une entreprise hasardeuse, car il ne permet pas de mesurer à quel point c'est bien le régime de reproduction qui est responsable de ce que l'on observe ou un artefact lié au locus étudié. Plus cette mesure est faite sur un grand nombre de marqueurs, plus fiables seront les inférences qu'on en tirera. La philosophie de pondération est la même que pour le F_{IS} multiallélique et ce sont donc les loci les plus polymorphes (qui ont le plus d'allèles aux fréquences les plus équilibrées) qui ont le plus de poids dans le calcul du F_{IS} moyen.

Enfin, il est plus fiable de calculer un F_{IS} moyen sur plusieurs réplicats indépendants (échantillons), la philosophie de pondération restant la même, additionné des tailles respectives des différents échantillons si celles-ci diffèrent. Il convient alors de définir le F_{IS} comme l'indice de fixation, ou degré relatif d'homozygotie des individus dans les sous-populations (d'où les lettres i et s en indice) provenant d'une rencontre non

aléatoire des allèles pour former les individus de chaque sous-population. La formule 6 devient (Nei et Chesser, 1983) :

$$F_{IS} = \frac{H_s - \overline{H_o}}{H_s} \qquad (9)$$

où H_s représente l'hétérozygotie attendue moyenne sur l'ensemble des sites, des loci et allèles ou, plus exactement, la diversité génétique moyenne sur l'ensemble des sous-échantillons, et $\overline{H_o}$ l'hétérozygotie moyenne observée. Cependant, afin de nous conformer aux notations et expressions modernes il nous faut maintenant exprimer cet indice en fonction des probabilités d'identité entre allèles. Soit Q_I la probabilité d'identité de deux allèles dans un individu à un locus prix au hasard et Q_S la probabilité d'identité de deux allèles pris au hasard dans deux individus de la même sous-population pour le même locus pris au hasard, alors nous avons Q_I = 1- $\overline{H_o}$ et Q_S = 1 - H_s et donc :

$$F_{IS} = \frac{1 - Q_S - 1 + Q_I}{1 - Q_S} = \frac{Q_I - Q_S}{1 - Q_S} \qquad (10)$$

Qui correspond à la définition la plus générale du F_{IS} (Rousset, 2004).

POPULATIONS STRUCTURÉES, EFFET WAHLUND ET STATISTIQUES *F* (F-STATISTICS)

L'exemple du modèle en îles

Les populations naturelles d'êtres vivants ne sont pas distribuées de façon homogène sur l'ensemble de la biosphère : elles sont subdivisées. Un très grand nombre de modèles de populations structurées existe. Le but de cette notice n'étant pas de passer en revue tout ce qui existe en génétique des populations (~ une dizaine de volumes de 500 pages chacun), nous nous focaliserons ici sur le modèle en îles de Wright (Wright, 1951). Nous allons supposer que la population qui nous intéresse est subdivisée en *n* sous-populations de taille *N* chacune, avec *n* très grand. À chaque génération, chaque population meurt en envoyant une infinité de propagules dans le milieu. Chaque sous-population est ensuite recolonisée par ces propagules avec une proportion *m* qui vient d'ailleurs et (1-*m*) qui revient à sa population d'origine (ils n'ont pas bougé en fait). Cela revient à dire que chaque sous-population est constituée, à chaque génération, de *Nm* immigrants et de (1-*m*)*N* résidents et où les immigrants proviennent de chacune des *n* sous-populations avec la même probabilité 1/*n* (elles ont toutes la même taille et les propagules tombent au hasard). Notons que cette probabilité est cependant faible (car *n* grand). Ce modèle est illustré dans la figure 7.

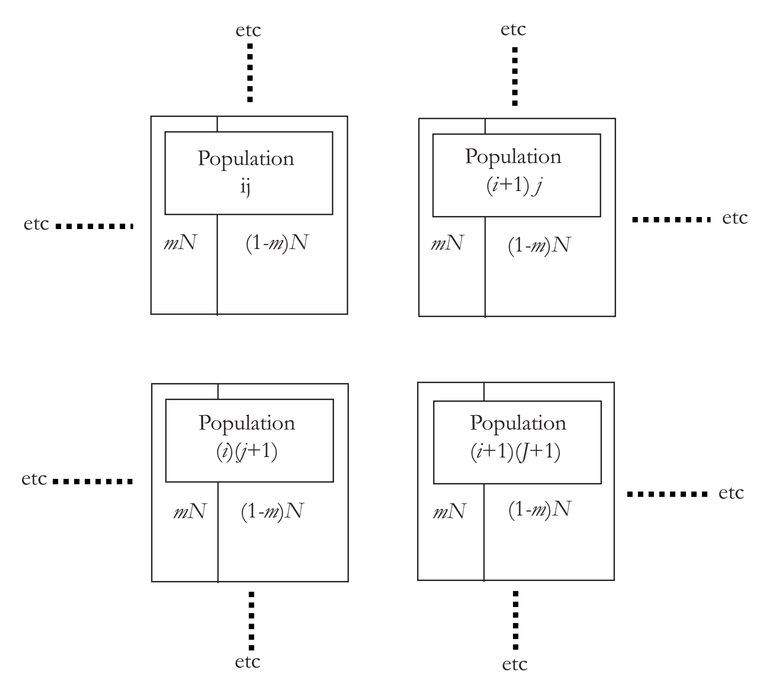


Figure 7
Le modèle en îles de Wright. Chacune des *n* sous-populations est constituée de *N* individus dont une proportion *m* provient de l'extérieur (migrants) et (1-*m*) d'autochtones.

Le déficit en hétérozygotes dû à la structuration (effet Wahlund)

Pour l'instant, on ne fait pas d'hypothèse sur le régime de reproduction, on va donc considérer que la reproduction est panmictique. Si on laisse ce système évoluer, les fréquences alléliques des différents loci vont donc évoluer également à l'intérieur des sous-populations, même si elles restent inchangées sur l'ensemble. Nous avons en effet supposé que n était très grand ($n \to \infty$). Il n'y a donc pas de dérive à l'échelle globale. Par contre, N et m sont limités, il y a donc possibilité de dérive génétique dans chaque sous-population, comme décrit en p. 31, et interaction avec la migration. La dérive va tendre à faire diverger les fréquences alléliques entre les différentes sous-populations et la migration va tendre à homogénéiser ces fréquences. Si on admet qu'il y a panmixie dans chaque sous-population i, on va observer, sur l'ensemble du système, une hétérozygotie de :

$$H_{oT} = \frac{1}{n}\sum_{i=1}^{i=n} 2p_i(1-p_i) \qquad (11)$$

s'il n'y a que deux allèles dans tout le système, dont le premier à la fréquence p_i dans la sous-population i.

Sur l'ensemble, la fréquence de cet allèle sera donc égale à la moyenne des fréquences trouvées sur l'ensemble des sous-populations :

$$\overline{p} = \frac{1}{n}\sum_{i=1}^{i=n} p_i \qquad (12)$$

Sur l'ensemble encore, on peut également calculer la fréquence des hétérozygotes attendus sous l'hypothèse d'une panmixie globale :

$$H_{eT} = 2\overline{p}\left(1-\overline{p}\right) \qquad (13)$$

On peut alors calculer un déficit global en hétérozygotes :

$$F_{IST} = \frac{H_{eT} - H_{oT}}{H_{eT}} \qquad (14)$$

En utilisant (11) et (13), on obtient pour (14) :

$$F_{IST} = \frac{2\overline{p}\left(1-\overline{p}\right) - \frac{1}{n}\sum_{i=1}^{n} 2p_i\left(1-p_i\right)}{2\overline{p}\left(1-\overline{p}\right)}$$

d'où

$$F_{IST} = \frac{2\left[\overline{p} - \overline{p}^2 - \frac{1}{n}\sum_{i=1}^{n}\left(p_i - p_i^2\right)\right]}{2\overline{p}\left(1-\overline{p}\right)}$$

en simplifiant et en développant :

$$F_{IST} = \frac{\overline{p} - \overline{p}^2 - \frac{1}{n}\sum_{i=1}^{n} p_i + \frac{1}{n}\sum_{i=1}^{n} p_i^2}{\overline{p}\left(1-\overline{p}\right)}$$

et donc en utilisant (12) :

$$F_{IST} = \frac{\overline{p} - \overline{p}^2 - \overline{p} + \overline{p^2}}{\overline{p}\left(1-\overline{p}\right)}$$

ce qui donne enfin :

$$F_{IST} = \frac{\overline{p^2} - \overline{p}^2}{\overline{p}\left(1-\overline{p}\right)} \qquad (15)$$

L'équation (15) peut également s'écrire (veuillez vous référer à la réponse 4 si vous ne voyez pas pourquoi) :

$$F_{IST} = \frac{\overline{\left(p_i - \overline{p}\right)^2}}{\overline{p}\left(1 - \overline{p}\right)}$$

Il s'agit donc du rapport entre la moyenne du carré des écarts à la moyenne (si cela ne vous rappelle rien, reportez-vous à la réponse 5) et la valeur que prend cette moyenne des carrés des écarts à la moyenne quand toutes les sous-populations sont fixées pour l'un ou l'autre des allèles (à vérifier en réponse 6). Dans le cas de deux allèles, cela veut dire qu'on a $\overline{p}$ sous-populations fixées pour l'allèle 1 et 1- $\overline{p}$ pour le 2. Nous avons donc :

$$F_{IST} = \frac{\sigma^2\left(p_i\right)}{\sigma^2_{\max}\left(p_i\right)} \qquad (16)$$

Cette valeur est nécessairement toujours positive et correspond donc à un déficit en hétérozygotes dû au fait que l'on calcule le F_{IS} en réunissant des individus qui n'appartiennent pas aux mêmes unités. On voit bien dans les équations (14), (15) et (16) que si les sous-populations partagent les mêmes fréquences alléliques (variance nulle), ce déficit est nul (pas de déviation par rapport à Hardy-Weinberg), alors que dans les autres cas il est positif, et ce d'autant plus que les fréquences alléliques diffèrent entre sous-populations, jusqu'à une valeur maximale de 1 quand chaque sous-population est fixée pour un des allèles présents (variance maximale). On appelle ce phénomène l'effet Wahlund (Wahlund, 1928), c'est-à-dire le déficit en hétérozygotes dû à la structuration de la population. Ce déficit en hétérozygotes correspond en fait au F_{ST} de Wright (1965), dont la formule en fonction des hétérozygoties et diversités géniques (Nei et Chesser, 1983) est la suivante :

$$F_{ST} = \frac{H_T - H_s}{H_T} \qquad (17)$$

où H_T correspond à l'hétérozygotie attendue si tous les individus de toutes les sous-populations se croisaient au hasard (panmixie globale) et H_s correspond à l'hétérozygotie moyenne attendue si les individus se croisaient au hasard à l'intérieur de chaque sous-population (panmixie locale). En fait pour le cas le plus général, H_T et H_s correspondent respectivement à la diversité génique de la population totale et à celle trouvée au sein des sous-populations (moyennée sur l'ensemble).

Les statistiques *F* de Wright (1965)

Définitions classiques

Il est possible que les sous-populations de notre modèle en îles ne soient pas panmictiques. Dans ce cas, le déficit en hétérozygotes global résultera de deux effets :

l'effet Wahlund et l'effet des croisements non aléatoires dans les sous-populations. On aura alors (Nei et Chesser, 1983) :

$$F_{IT} = \frac{H_T - \overline{H_o}}{H_T} \quad (18)$$

Nous pouvons ainsi définir les trois statistiques *F* de Wright (ou indices de fixation de Wright). Le F_{IS} (I pour individu et S pour sous-population) mesure la consanguinité des individus eux-mêmes relativement à la consanguinité entre individus d'une même sous-population (parenté). C'est aussi une mesure de la part d'homozygotie qui provient d'une déviation par rapport au régime de reproduction panmictique idéal dans les sous-populations (rencontre au hasard des gamètes dans chaque sous-population), on dit souvent aussi que le F_{IS} mesure le déficit en hétérozygotes local moyen (sur l'ensemble des sous-populations). Le F_{ST} correspond à la consanguinité entre individus d'une même sous-population relativement à la consanguinité entre sous-populations de la population totale. Il mesure l'effet Wahlund (ou structuration des populations), c'est-à-dire la part d'homozygotie des individus de la population totale (d'où l'indice T) provenant de la subdivision de ces derniers en sous-populations de tailles limitées (indice S), on dit aussi qu'il mesure la différenciation génétique entre sous-populations. Enfin, le F_{IT} mesure l'homozygotie des individus de la population totale résultant des deux phénomènes précédents :

$$\begin{cases} F_{IS} = \dfrac{H_s - H_o}{H_s} \\ F_{ST} = \dfrac{H_T - H_s}{H_T} \\ F_{IT} = \dfrac{H_T - H_o}{H_T} \end{cases} \quad (19)$$

À partir des équations (19), il est facile d'obtenir la relation classique (au moins pour les personnes ayant déjà entendu parler de génétique des populations structurées) :

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \quad (20)$$

Il n'est pas inintéressant de préciser que ces indices de fixation mesurent également, à un certain degré, la consanguinité des individus, celle due au nombre restreint de partenaires dans des sous-populations isolées et de tailles finies (F_{ST}) et celle due aux déviations par rapport à un régime panmictique de reproduction (F_{IS}).

Comme nous l'avons vu, le F_{IS} varie de – 1 à 1 (voir p. 41). Le F_{ST} varie de 0 (pas de structuration) à 1 (toutes les sous-populations sont fixées pour l'un ou l'autre des allèles). Le F_{IT}, tout comme le F_{IS}, varie entre – 1 et 1.

Nous pouvons, pour finir, remarquer que, pour un modèle en îles infini et deux allèles, nous avons démontré avec l'équation (16) que :

$$F_{ST} = \frac{\sigma^2(p)}{\sigma^2_{\max}(p)}$$

ce qui correspond à la définition originelle du F_{ST} (Wright, 1965) restée assez populaire pour qu'on la trouve encore dans de nombreuses publications et ouvrages spécialisés.

Définitions en fonction des probabilités d'identité

Maintenant, notons Q_I la probabilité d'identité de deux allèles d'un même locus au sein d'un même individu pris au hasard, Q_S la probabilité de tirer deux allèles identiques d'un même locus de deux individus différents de la même sous-population et Q_T la probabilité de tirer deux allèles identiques de deux individus de deux sous-populations différentes pris au hasard. Nous pouvons alors donner les formules généralisées des statistiques F pour un degré 3 de subdivision (individu, sous-population et total) (Rousset, 2004) :

$$\begin{cases} F_{IS} = \dfrac{Q_I - Q_S}{1 - Q_S} \\ F_{ST} = \dfrac{Q_S - Q_T}{1 - Q_T} \\ F_{IT} = \dfrac{Q_I - Q_T}{1 - Q_T} \end{cases} \quad (21)$$

En partant du système d'équations (21), nous pouvons également constater que le F_{ST} varie entre $F_{ST} = 0$, quand l'identité génétique entre individus est indépendante de la sous-population où ces individus résident (absence de différenciation génétique entre sous-populations), et $F_{ST} = 1$, quand tous les individus résidant dans la même sous-population sont génétiquement identiques ($Q_S = 1$), mais ne sont pas nécessairement identiques à ceux résidant dans d'autres sous-populations ($Q_T < 1$). Par conséquent, $F_{ST} = 1$ signifie une indépendance complète des sous-populations (et donc des individus qui les composent) entre elles, ce qui est attendu si ces sous-populations sont restées isolées les unes des autres pendant une durée suffisamment longue. Le F_{IT} varie entre $F_{IT} = -1$, quand tous les individus de la population totale sont hétérozygotes pour les deux mêmes allèles et $F_{IT} = 1$ quand tous les individus sont homozygotes avec au moins deux allèles dans la population totale.

Quand la probabilité d'échantillonner deux allèles identiques d'un même locus sur l'ensemble de la métapopulation devient indépendante de la localité d'origine et des individus d'où l'on peut les tirer, alors $Q_I = Q_S = Q_T$ et une conformité globale aux proportions attendues sous Hardy-Weinberg est observée avec $F_{IS} = F_{ST} = F_{IT} = 0$.

Inférer Nm à partir du F_{ST} dans un modèle en îles

Nous avons vu précédemment qu'en utilisant les conséquences analytiques de l'autofécondation, nous pouvions estimer un taux possible d'autofécondation à partir de la connaissance du F_{IS} (équation 8). Nous allons voir maintenant que la connaissance d'un F_{ST} peut permettre l'inférence du nombre d'individus migrants (le produit Nm) dans une sous-population si cette dernière fait partie d'un modèle en îles. Dans un modèle en îles infini composé de sous-populations panmictiques, la probabilité d'identité entre deux allèles pris au hasard entre deux sous-populations est nulle. En effet, si le nombre de sous-populations n est suffisamment grand, cette probabilité est égale à $Q_T = 1/n \approx 0$. Ceci conduit naturellement à ce que $F_{ST} = Q_S$, la probabilité d'identité entre allèles d'individus résidant dans la même sous-population (voir l'équation 21). Soit $Q_{S(t)}$ cette probabilité à une génération quelconque t. La proportion d'allèles non identiques dans chaque sous-population est donc égale à $(1 - Q_{S(t)})$. À t+1, la proportion d'allèles identiques se verra augmentée par les allèles échantillonnés deux fois parmi ceux différents au temps t. Sachant que la probabilité d'échantillonner deux fois le même allèle parmi les $2N$ existants est égale à $(1/2N)^2$, et qu'il faut répéter l'opération $2N$ fois pour construire une sous-population, on a donc $1/2N$ chances de prélever deux fois le même allèle parmi les $(1 - Q_{S(t)})$ qui diffèrent au temps t. L'accroissement de la probabilité d'identité dans les sous-populations sera donc de $(1 - Q_{S(t)})/2N$ et, si on ignore la migration, nous aurons $Q_{S(t+1)} = Q_{S(t)} + (1 - Q_{S(t)})/2N$. Avec la migration, cette probabilité ne reste valable que pour les paires d'allèles non migrants, avec la probabilité $(1 - m)^2$, car les immigrants ne peuvent être identiques à personne ($Q_T \approx 0$). En tenant compte de l'ensemble de ces informations, et en espérant que les lecteurs ne sont pas encore entièrement perdus, nous pouvons poser qu'à la génération t+1 :

$$Q_{S(t+1)} = (1-m)^2 \left[Q_{S(t)} + (1 - Q_{S(t)}) \frac{1}{2N} \right] \tag{22}$$

À l'équilibre entre migration et dérive, nous aurons :

$$Q_{S(t+1)} = Q_{S(t)} = \hat{Q}_S = \frac{\frac{(1-m)^2}{2N}}{1 - (1-m)^2 + \frac{(1-m)^2}{2N}}$$

ce qui donne :

$$\hat{Q}_S = \frac{(1-m)^2}{2Nm(2-m) + 1 - 2m + m^2}$$

Ici, nous pouvons supposer que m est suffisamment petit pour pouvoir le négliger en face de 1. Si alors nous remplaçons Q_S par F_{ST}, nous pouvons retrouver la formule classique :

$$F_{ST} \approx \frac{1}{4Nm + 1} \tag{23}$$

De cette équation, il est facile d'extraire la non moins fameuse égalité $Nm = (1 - F_{ST})/4F_{ST}$.

Nous pouvons introduire ici le fait que si les allèles sont affectés par un taux de mutation constant u, correspondant au modèle IAM, alors l'équation (23) devient :

$$F_{ST} \approx \frac{1}{4N(m+u)+1} \quad (24)$$

En partant de l'équation (24), il est très facile de constater que pour des loci soumis à de forts taux de mutation et donc très polymorphes, la valeur maximale du F_{ST} ne pourra jamais atteindre la valeur 1, même quand $m = 0$. Cela signifie également qu'un petit F_{ST} peut être obtenu dans des populations très structurées (Nm petits) si les loci utilisés sont très polymorphes (beaucoup d'allèles, taux de mutation élevé). Il existe une méthode simple pour se rendre compte de ce phénomène, comme nous le verrons plus loin (en p. 60).

Pour obtenir l'équation (24), nous avons également fait l'hypothèse d'un nombre infini d'allèles possible (IAM). Il se peut cependant que le modèle de mutation s'écarte fortement de cet idéal, ce qui peut altérer les valeurs d'équilibre des statistiques F (Rousset, 1996). Dans le cas d'un KAM, Rousset (1996) montre que les statistiques F attendues sont les mêmes que pour un IAM, mais avec un taux de mutation augmenté de $K/(K - 1)$. Pour l'équation (24), cela donnerait :

$$F_{ST} \approx \frac{1}{4N(m+\frac{K}{K-1}u)+1} \quad (25)$$

D'une manière très analogue, dans le cas d'un modèle en îles fini (nombre d'îles n petit), on montre que l'équation (23) devient (toujours pour m petit) (en partant par exemple de Rousset, 1996) :

$$F_{ST} \approx \frac{1}{4Nm\frac{n}{n-1}+1} \quad (26)$$

Il est également utile de remarquer que les équations (23) et (24) sont obtenues sous l'hypothèse d'un modèle en îles infini à l'équilibre entre migration, mutation et dérive. Relâcher ces hypothèses peut alors fortement limiter nos capacités d'inférences sur le nombre effectif de migrants (Whitlock et McCauley, 1998). Ainsi, l'estimation du Nm à partir du F_{ST} n'est bien souvent qu'un « équivalent modèle en îles ». C'est aussi pour ce genre de problèmes que d'autres types de modèles de populations structurées ont été imaginés.

Pertinence du modèle en îles

Le modèle en îles n'a pas que l'avantage de simplifier les analyses mathématiques. Ce modèle est en effet conforme, même approximativement, à certaines structures de

populations réelles. Dans le milieu marin, par exemple, il est probable que ce modèle reflète le cycle de nombreux types d'organismes fixés et à dispersion importante comme les bivalves, les échinodermes, les crustacés ou les algues, mais aussi de nombreux parasites tels que les crustacés parasites (copépodes, cirripèdes, isopodes, amphipodes) et autres monogènes (De Meeûs, 2000), ou même les parasites en général si on considère l'individu hôte comme une sous-population et que ces individus hôtes sont suffisamment mobiles (Nébavi *et al.*, 2006). Malgré un aspect très caricatural, le modèle en îles représente donc un outil souvent efficace pour étudier les populations naturelles, notamment de parasites.

Autres modèles de populations structurées

Il existe d'autres modèles de populations structurées qui permettent d'étudier les conséquences génétiques d'autres contraintes de subdivision que celles décrites par le modèle en îles. Ces modèles font intervenir une composante géographique où l'éloignement des sous-populations et/ou des individus va influencer les probabilités d'échanges de gènes et/ou d'individus (flux de gènes et/ou d'individus). En termes de migration formelle, ceci peut se traduire par un schéma discontinu de migration comme pour les modèles en pas japonais (*stepping stone models*) (Kimura et Weiss, 1964 ; Slatkin, 1985). Le modèle de diffusion peut être continu dans le cas de modèles en voisinage (*neighbourhood models*) (Wright, 1965 ; Rousset, 2000 ; Leblois *et al.*, 2004).

Ces modèles de populations peuvent se présenter en une dimension, comme dans le cas d'espèces inféodées à un écotone bien défini (écosystèmes côtiers, bordures de chemins, de routes de forêts, etc.), deux dimensions (paysage quelconque) ou trois dimensions (milieux aquatiques, forestiers, etc.). La figure 8 illustre ces différentes possibilités pour un modèle en *stepping stone*. Dans cette figure, les migrants ne peuvent passer que d'une population directement adjacente à l'autre. Il existe également des modèles mixtes entre *stepping stone* et modèle en îles (voir Hartl et Clark, 1989 : 317-318). Enfin, le problème des individus ou sous-populations marginales (en situation de bordure) est souvent résolu par la connexion entre elles de ces bords libres, aboutissant à l'établissement d'un cercle (modèles en une dimension) ou d'un tore (modèle en deux dimensions).

Dans de telles configurations de populations, plutôt que d'étudier un F_{ST} global, il est plus informatif d'examiner la corrélation qui relie les distances génétiques avec les distances géographiques séparant les paires d'individus ou de dèmes (Rousset, 1997, 2000).

Estimateurs non biaisés des statistiques *F*

Les définitions présentées dans les équations (19) et (21) correspondent aux définitions paramétriques des *F* de Wright. Dans la réalité, le nombre de sous-populations et le

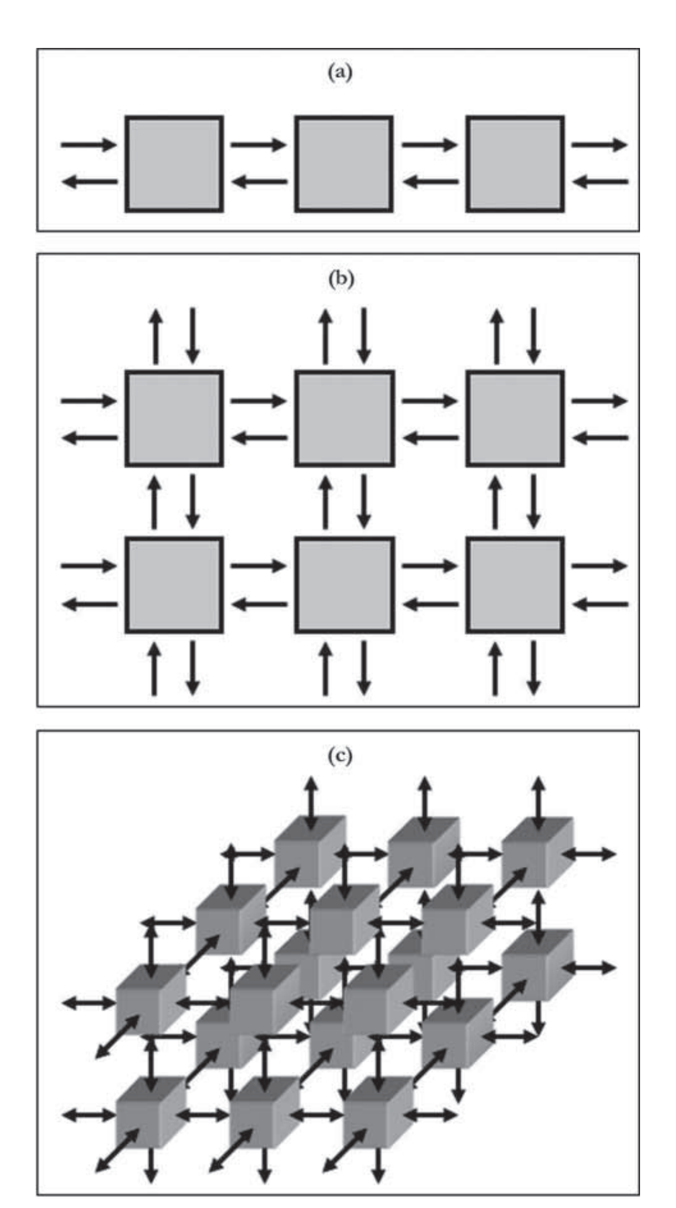


Figure 8
Représentation graphique des modèles en pas japonais (*stepping stone*) à une (a), deux (b) et trois (c) dimensions. Dans ces modèles, chaque dème de taille *N* n'échange ses *Nm* migrants qu'avec les dèmes adjacents.

nombre d'individus échantillonnés par sous-population sont tous les deux limités. Le génotypage d'individus sur plusieurs marqueurs génétiques ne peut se faire que sur quelques sous-populations et sur un échantillon d'individus de ces sous-populations.

Nous pouvons ici faire un petit rappel de statistiques de base. Pour un échantillon de taille *n* où on mesure un caractère variable *x* dont la moyenne est $\overline{x}$, la variance aura la forme :

$$\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (27)$$

si et seulement si on a échantillonné et mesuré *x* sur tous les individus de la population. On aura alors une mesure paramétrique de la variance.

Dans le cas contraire, on doit appliquer la formule d'estimation de cette variance à partir de notre échantillon de taille *n* :

$$s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (28)$$

Ceci vient du fait que pour calculer la moyenne, nous avons utilisé toute l'information concernant la somme des x_i. Par conséquent, quand on procède à la somme des $(x_i - \overline{x})^2$ et que l'on arrive au $(n\text{-}1)^{\text{ème}}$ terme, on a déjà toute l'information, le $n^{\text{ème}}$ terme apporte une information redondante. C'est pour cela que l'on divise par *n*-1 (degré de liberté). Si on divise par *n* au lieu de (*n*-1), on sous-estime la variance (estimation biaisée). On voit bien que plus *n* augmente, moins le biais est important.

Pour les *F* de Wright, que l'on peut faire correspondre à des rapports de variance, le même type de phénomène se rencontre. Imaginons par exemple que je suis très paresseux et que je n'échantillonne qu'un individu par sous-population. Dans ce cas, j'aurais un individu soit homozygote, soit hétérozygote dans chacune de mes sous-populations. Je ne peux calculer alors un F_{IS} que dans les sites polymorphes, c'est-à-dire ceux où j'ai trouvé un hétérozygote, avec $p_1 = p_2 = 0{,}5$. En appliquant l'équation (19) on calcule :

$$F_{IS} = \frac{H_s - H_o}{H_s} = \frac{2p_1p_2 - 1}{2p_1p_2} = -1$$

Le biais est ici énorme : on sous-estime le déficit de 100 % puisque, en effet, on ne peut s'attendre à rien d'autre que cette valeur de - 1, qui bien évidemment n'a pas d'autre sens.

L'estimation non biaisée des paramètres *F* est beaucoup plus complexe que pour une simple variance. Les estimateurs *f*, θ et *F* de Weir et Cockerham (Weir et Cockerham, 1984) sont des estimateurs non biaisés des F_{IS}, le F_{ST} et le F_{IT} de Wright respectivement. Ils sont issus d'un modèle d'analyse de variance hiérarchique (*nested analysis of variance*) des fréquences alléliques dans les individus des sous-populations, entre individus des sous-populations et entre sous-populations. En reprenant les notations originales, les estimateurs de Weir et Cockerham dépendent donc de σ_a^2, σ_b^2 et σ_w^2 qui sont les composantes inter dèmes (*among sub-populations*, *a*), entre individus de chaque sous-population (*between individuals*, *b*) et intra-

individuelle (*within individuals*, *w*) de la variance des fréquences alléliques. À partir de là, on peut exprimer les différents estimateurs sous la forme :

$$\begin{cases} f = \dfrac{\sigma_b^2}{(\sigma_b^2 + \sigma_w^2)} \\ \theta = \dfrac{\sigma_a^2}{(\sigma_a^2 + \sigma_b^2 + \sigma_w^2)} \\ F = \dfrac{\sigma_a^2 + \sigma_b^2}{(\sigma_a^2 + \sigma_b^2 + \sigma_w^2)} \end{cases} \quad (29)$$

Le calcul de ces composantes s'effectue à partir de la table d'analyse de variance hiérarchique similaire à celle présentée dans le tableau 2.

Tableau 2
Analyse de variance des fréquences alléliques dans un échantillon subdivisé en *n* sous-échantillons tous de taille N chacun.

Source de variation	Ddl	*MS* observés	*MS* attendus
Entre sous-échantillons (*a*)	n - 1	MS_a	$2N\ \sigma_a^2 + 2\ \sigma_b^2 + \sigma_w^2$
Entre individus dans chaque sous-échantillon (*b*)	$n(N - 1)$	MS_b	$2\ \sigma_b^2 + \sigma_w^2$
Entre allèles dans chaque individu (*w*)	$nN(2 - 1) = N_T$	MS_w	σ_w^2

Ddl : Degré de liberté ; *MS* : Carrés moyens attendus (*Mean Squares*).

Il s'agit ensuite de calculer les carrés moyens des fréquences d'allèles de l'échantillon pour obtenir, avec les carrés moyens attendus, un système de trois équations à trois inconnues permettant de déduire les différentes composantes de la variance des fréquences alléliques. Cela est possible en s'aidant d'un ouvrage traitant en détail de l'analyse de variance hiérarchique (*nested* en anglais) (Sokal et Rohlf, 1981). Si nous considérons le cas d'un locus à deux allèles (1 et 2), l'étude de la variation allélique se fait alors sur un seul allèle (l'allèle 1) qui prend la valeur α = 1 ou α = 0 quand il est présent ou absent. Nous avons besoin ensuite de calculer les sommes des carrés des quantités de l'allèle dans chaque chromosome de chaque individu (SS_1), des individus (SS_2), des sous-échantillons (SS_3) et de l'échantillon total (SS_4). Notons N_{T11} le nombre d'homozygotes pour l'allèle 1, N_{T12} celui des hétérozygotes et N_{T22} celui des homozygotes pour l'absence de cet allèle dans l'échantillon total. De même, considérons N_{i11} le nombre d'homozygotes pour l'allèle 1, N_{i12} celui des hétérozygotes et N_{i22} celui des homozygotes pour l'absence de cet allèle dans l'échantillon *i*. Sachant que la quantité mesurée α prend la valeur 0 ou 1 et en supposant

une espérance uniforme des hétérozygoties de chaque sous-population, nous pouvons poser[3] :

$$\begin{cases} SS_1 = \sum_1^n \sum_1^N \sum_1^2 \alpha^2 \\ SS_2 = \dfrac{\sum_1^n \sum_1^N \left(\sum_1^2 \alpha \right)^2}{2} \\ SS_3 = \dfrac{\sum_1^n \left(\sum_1^N \sum_1^2 \alpha \right)^2}{2N} \\ SS_4 = \dfrac{\left(\sum_1^n \sum_1^N \sum_1^2 \alpha \right)^2}{2N_T} \end{cases} \quad (30)$$

Du système d'équations (30), on peut tirer (en détaillant au maximum) :

$$\begin{cases} SS_1 = N_{T11}\left[2\left(1^2\right)\right] + N_{T12}\left[1\left(1^2\right)\right] + N_{T22}\left[2\left(0^2\right)\right] = 2N_{T11} + N_{T12} \\ SS_2 = \dfrac{N_{T11}(2)^2 + N_{T12}(1)^2 + N_{T22}(0)^2}{2} = 2N_{T11} + \dfrac{N_{T12}}{2} \\ SS_3 = \dfrac{\sum_1^n \{N_{i11}[2(1)] + N_{i12}[1(1)]\}^2}{2N} = \dfrac{\sum_1^n \{2N_i p_i\}^2}{2N} = \dfrac{4N^2 \sum_1^n p_i^2}{2N} \\ SS_4 = \dfrac{\{N_{T11}[2(1)] + N_{T12}[1(1)] + N_{T22}[2(0)]\}^2}{2N_T} = \dfrac{\{2N_{T11} + N_{T12}\}^2}{2N_T} \end{cases} \quad (31)$$

Du système (31) on déduit :

$$\begin{cases} SS_1 = 2N_T \overline{p} \\ SS_2 = 2N_T \overline{p} - \dfrac{N_{T12}}{2} \\ SS_3 = 2nN \dfrac{1}{n} \sum_1^n p_i^2 = 2N_T \overline{p^2} \\ SS_4 = \dfrac{\{2N_T \overline{p}\}^2}{2N_T} = 2N_T \overline{p}^2 \end{cases} \quad (32)$$

[3] Le détail de cette démonstration n'est pas indispensable, mais je conseille à tous les lecteurs d'essayer de la comprendre au moins une fois.

À partir de ces sommes de carrés, nous pouvons ensuite calculer les composantes intra-individuelles (*w*), interindividuelles (*b*) et inter-sous-échantillons (*a*) des sommes de carrés de cette quantité α, soit SS_w, SS_b et SS_a respectivement :

$$\begin{cases} SS_w = SS_1 - SS_2 \\ SS_b = SS_2 - SS_3 \\ SS_a = SS_3 - SS_4 \end{cases} \quad (33)$$

c'est-à-dire les contributions respectives de ces différents niveaux à la variance des fréquences alléliques. Des systèmes d'équations (33) et (32), on peut obtenir :

$$\begin{cases} SS_w = 2N_T\overline{p} - 2N_T\overline{p} + \dfrac{N_{T12}}{2} \\ SS_b = 2N_T\overline{p} - \dfrac{N_{T12}}{2} - 2N_T\overline{p^2} = 2N_T\left(\overline{p} - \overline{p^2}\right) - \dfrac{N_{T12}}{2} \\ SS_a = 2N_T\overline{p^2} - 2N_T\overline{p}^2 = 2N_T\left(\overline{p^2} - \overline{p}^2\right) \end{cases} \quad (34)$$

Toujours en détaillant sensiblement, nous déduisons du système d'équations (34) :

$$\begin{cases} SS_w = \dfrac{N_{T12}}{2} \\ SS_b = 2N_T\left(\overline{p} - \overline{p^2} + \overline{p}^2 - \overline{p}^2\right) - \dfrac{N_{T12}}{2} = 2N_T\left(\overline{p}(1-\overline{p}) - \sigma^2(p)\right) - \dfrac{N_{T12}}{2} \\ SS_a = 2N_T\overline{p^2} - 2N_T\overline{p}^2 = 2N_T\left(\overline{p^2} - \overline{p}^2\right) = 2N_T\sigma^2(p) \end{cases} \quad (35)$$

d'où on tire :

$$\begin{cases} SS_w = \dfrac{N_{T12}}{2} \\ SS_b = 2N_T\left(\sigma^2_{\max}(p) - \sigma^2(p)\right) - \dfrac{N_{T12}}{2} \\ SS_a = 2N_T\sigma^2(p) \end{cases} \quad (36)$$

Nous pouvons enfin obtenir les moyennes de ces sommes de carrés en les divisant par leur degré de liberté respectif et ainsi obtenir le système d'équations à trois inconnues :

$$\begin{cases} MS_w = \dfrac{\frac{N_{T12}}{2}}{N_T} = \dfrac{N_{T12}}{2N_T} = \sigma_w^2 \\ MS_b = \dfrac{2N_T\left(\sigma_{\max}^2(p) - \sigma^2(p)\right) - \frac{N_{T12}}{2}}{n(N-1))} = 2\sigma_b^2 + \sigma_w^2 \\ MS_a = \dfrac{2N_T\sigma^2(p)}{n-1} = 2N\sigma_a^2 + 2\sigma_b^2 + \sigma_w^2 \end{cases} \quad (37)$$

On a donc :

$$\begin{cases} \sigma_w^2 = \dfrac{N_{T12}}{2N_T} \\ \sigma_b^2 = \dfrac{2N_T\left(\sigma_{\max}^2(p) - \sigma^2(p)\right) - \frac{N_{T12}}{2}}{2n(N-1))} - \dfrac{N_{T12}}{4N_T} \\ \sigma_a^2 = \dfrac{N_T\sigma^2(p)}{N(n-1)} - \dfrac{2N_T\left(\sigma_{\max}^2(p) - \sigma^2(p)\right) - \frac{N_{T12}}{2}}{2Nn(N-1))} \end{cases} \quad (38)$$

ce qui donne :

$$\begin{cases} \sigma_w^2 = \dfrac{N_{T12}}{2N_T} \\ \sigma_b^2 = \dfrac{N\left(\sigma_{\max}^2(p) - \sigma^2(p)\right)}{N-1} - \dfrac{N_{T12}}{4n(N-1))} - \dfrac{N_{T12}}{4N_T} \\ \sigma_a^2 = \dfrac{n\sigma^2(p)}{(n-1)} - \dfrac{\sigma_{\max}^2(p) - \sigma^2(p)}{N-1} - \dfrac{N_{T12}}{4N_T(N-1)} \end{cases} \quad (39)$$

La combinaison des systèmes d'équations (39) et (29) permet d'obtenir les estimateurs des différentes statistiques F dans le cas de figure présenté.

Ceux qui souhaiteraient plus de détails sont invités à consulter la bibliographie correspondante, car je n'entrerai pas plus dans les détails ici étant donné que ces estimateurs sont calculés par la plupart des logiciels disponibles tels que Fstat 2.9.3 (Goudet, 2002) téléchargeable gratuitement à http://www.unil.ch/izea/softwares/fstat.html (voir Goudet, 1995), Genetix 4.03 (Belkhir *et al.*, 2004) téléchargeable gratuitement à http://www.univ-montp2.fr/~genetix/genetix/genetix.htm, ou encore Genepop 3.4 (Raymond et Rousset, 2003) téléchargeable gratuitement à http://wbiomed.curtin.edu.au/genepop/ (voir Raymond et Rousset, 1995b), Genepop 4 (Rousset, 2008) (http://kimura.univ-montp2.

fr/~rousset/Genepop.htm) ou la version web du logiciel (http://genepop.curtin.edu.au/).

Il est cependant nécessaire de signaler que si f et F varient entre – 1 et 1, tout comme les paramètres qu'ils estiment F_{IS} et F_{IT}, θ, quant à lui, varie entre – 1 et 1, alors que le paramètre qu'il estime, F_{ST} varie entre 0 et 1. L'estimateur du F_{ST} peut prendre des valeurs négatives, car sous l'hypothèse d'absence de structuration, θ, puisqu'il est non biaisé, doit être centré sur 0, valeur attendue sous l'hypothèse d'absence de structuration génétique. Des valeurs très négatives de θ correspondront à des configurations particulières où les sous-échantillons sont plus proches génétiquement que ce qui est attendu par variance d'échantillonnage. En effet, si on échantillonne deux fois dans la même population, on aura peu de chances d'échantillonner exactement le même nombre d'individus de chaque génotype dans chacun des deux échantillons. Il s'ensuit une différence aléatoire (variance d'échantillonnage) prise en compte par θ, dont l'espérance mathématique est centrée sur 0. Il est enfin utile de préciser que, pour plus de deux allèles, l'estimateur moyen pondère par construction les valeurs obtenues pour chaque allèle par le facteur $\overline{p}(1-\overline{p})$ (variance maximale possible dans l'équation 39), ce qui donne le maximum de poids aux allèles dont la fréquence est la moins proche de 0 et 1. Étant donné que les estimateurs moyens sur plusieurs loci et/ou plusieurs sous-échantillons suivent la même logique, ce sont les loci et/ou les sous-échantillons les plus polymorphes qui auront le plus de poids. D'autres méthodes de pondération existent. En particulier, celle proposée par Robertson et Hill (1984) a connu un certain succès pour ses propriétés statistiques (Goudet *et al.*, 1996 ; Rousset et Raymond, 1995). Ici, une pondération différente est implémentée. Si les estimateurs de Weir et Cockerham et Robertson et Hill sont notés avec les indices WC et RH respectivement, nous obtenons pour K allèles noté de A = 1 à K :

$$\begin{cases} f_{WC} = \dfrac{\sum_{A=1}^{A=K} \sigma_b^2(A)}{\sum_{A=1}^{A=K} \left[\sigma_b^2(A) + \sigma_w^2(A)\right]} \\ \theta_{WC} = \dfrac{\sum_{A=1}^{A=K} \sigma_a^2(A)}{\sum_{A=1}^{A=K} \left[\sigma_a^2(A) + \sigma_b^2(A) + \sigma_w^2(A)\right]} \\ F_{WC} = \dfrac{\sigma_a^2(A) + \sigma_b^2(A)}{\sum_{A=1}^{A=K} \left[\sigma_a^2(A) + \sigma_b^2(A) + \sigma_w^2(A)\right]} \end{cases} \quad (40)$$

pour les estimateurs de Weir et Cockerham où nous savons que les termes contiennent une pondération inhérente donnant davantage de poids aux allèles dont la fréquence est la plus proche de 0,5, et :

$$\begin{cases} f_{RH} = \dfrac{1}{K-1}\displaystyle\sum_{A=1}^{A=K}\dfrac{(1-p_A)\sigma_b^2(A)}{\left[\sigma_b^2(A)+\sigma_w^2(A)\right]} \\ \\ \theta_{RH} = \dfrac{1}{K-1}\displaystyle\sum_{A=1}^{A=K}\dfrac{(1-p_A)\sigma_a^2(A)}{\left[\sigma_a^2(A)+\sigma_b^2(A)+\sigma_w^2(A)\right]} \\ \\ F_{RH} = \dfrac{1}{K-1}\displaystyle\sum_{A=1}^{A=K}\dfrac{(1-p_A)\left[\sigma_a^2(A)+\sigma_b^2(A)\right]}{\left[\sigma_a^2(A)+\sigma_b^2(A)+\sigma_w^2(A)\right]} \end{cases} \quad (41)$$

pour les estimateurs de Robertson et Hill qui donnent un poids maximal aux allèles les plus rares (pondération par 1 - p_A). Les estimateurs de Weir et Cockerham sont non biaisés, mais sujets à une variance importante, alors que ceux de Robertson et Hill sont biaisés, mais beaucoup moins variables pour de faibles valeurs des F (Rousset et Raymond, 1995 ; Raufaste et Bonhomme, 2000), ce qui leur confère un avantage statistique certain (voir plus loin).

Les estimations multilocus tiennent également compte du polymorphisme des loci (les plus polymorphes auront en principe le plus de poids) et du nombre d'individus génotypés (par toujours le même nombre par locus), de même que les estimations multi-échantillons (pour le F_{IS}).

Mesures de différenciation génétique alternatives au F_{ST}

Les R-Statistiques

Dans le cas des microsatellites, si la mutation suit strictement un SMM (voir p. 32), il peut alors être plus approprié d'utiliser des mesures qui tiennent compte de la taille des allèles. Pour évaluer la différenciation entre sous-populations, Slatkin (1995) a proposé le R_{ST} dont la mesure tient compte de la taille des allèles, des allèles de taille proche ayant plus de chances d'avoir un ancêtre commun proche. Ces statistiques sont estimées d'une façon équivalente aux estimateurs de Weir et Cockerham (1984), sauf que ce sont les tailles des allèles et non leurs fréquences qui sont utilisées (Slatkin, 1995 ; Rousset, 1996 ; Michalakis et Excoffier, 1996). Le même principe peut être appliqué au F_{IS} (Rousset, 1996). Ces statistiques s'avèrent peu appropriées si le modèle de mutation dévie un peu du schéma idéal d'un SMM et sont de toutes manières sujettes à de trop fortes variances d'estimation. En règle générale, on préfère utiliser les estimateurs de Weir et Cockerham (1984) (Balloux *et al.*, 2000 ; Balloux et Goudet, 2002).

Le F_{ST} maximum possible

Dans le cas de marqueurs génétiques hypervariables comme les microsatellites, la valeur maximale du F_{ST} ne sera pas 1, car il y a plus d'allèles que de sous-échantillons (voir l'équation 24). Donc même si aucun sous-échantillon n'a d'allèle en commun (différenciation maximale possible), le F_{ST} pourra être de valeur modeste (Hedrick, 1999). Pour pallier ce problème, Hedrick (1999, 2005) propose une méthode simple pour visualiser de combien le F_{ST} observé est éloigné de sa valeur maximale que l'on observerait sans migration entre sous-populations. Dans une telle situation, et si le nombre de dèmes est assez grand, nous savons par l'équation (21) que $Q_T = 0$ (probabilité d'identité entre individus de dèmes différents) et que le F_{ST} est alors égal à $Q_s = 1 - H_s = F_{ST\max}$, où H_s est l'estimateur non biaisé de la diversité génétique de Nei (Nei et Chesser, 1983). On peut ensuite diviser la valeur observée dans les données par cette valeur maximale afin d'avoir une meilleure appréciation (et non pas une mesure exacte) du flux de gènes échangé entre les sous-populations échantillonnées $F_{ST}' = F_{ST}/F_{ST\max}$. Une alternative pour calculer ce $F_{ST\max}$ consiste en un recodage des allèles de telle sorte que les diversités locales restent les mêmes dans chaque sous-échantillon, mais aucun allèle en commun n'est partagé et chaque sous-échantillon montre des allèles uniquement présents chez lui (Meirmans, 2006). De mon expérience, les deux méthodes donnent des résultats très proches.

Différenciation génétique par paire d'échantillons ou d'individus

Il existe fréquemment des situations où la différenciation génétique doit être appréciée entre paires de populations ou même d'individus. Plusieurs possibilités s'offrent à nous. Le F_{ST} peut bien entendu être utilisé, mais il a été montré que dans cette configuration, il est loin d'être le plus performant (Rousset, 1997 ; Balloux et Goudet, 2002). L'empiriste avisé préférera l'utilisation d'autres outils, à choisir en fonction de la question posée. Si un isolement par la distance est recherché, l'utilisation de $F_{ST}/(1 - F_{ST})$ ou son estimateur $\theta/(1 - \theta)$ est recommandée par Rousset (1997). Nous verrons plus loin que ce nouvel estimateur est surtout utile pour inférer les paramètres démographiques de la population investiguée. Dans les autres situations, la distance harmonique (*chord distance*) de Cavalli-Sforza et Edwards (Cavalli-Sforza et Edwards, 1967) donne de meilleurs résultats (Takezaki et Nei, 1996 ; Kalinowski, 2002). Cette distance est obtenue suivant la formule suivante :

$$D_c = \frac{2}{r\pi}\sum_{j=1}^{r}\sqrt{2\left[1-\sum_{i=1}^{mj}\sqrt{x_{ij}y_{ij}}\right]} \qquad (42)$$

où r correspond au nombre de loci, j au label du locus (de 1 à r), i au label de l'allèle (de 1 à mj), mj au nombre d'allèles au locus j, x_{ij} et y_{ij} les fréquences de l'allèle i au locus j pour les sous-populations x et y respectivement.

Quand c'est la distance génétique entre individus qui est pertinente, il semble plus approprié d'utiliser la distance d'allèles partagés (*shared allelic distance*) (Bowcock *et al.*, 1994) (voir Prugnolle *et al.*, 2005). Si N_{sa} correspond au nombre d'allèles en commun partagés par deux individus sur l'ensemble des L loci, alors cette distance est égale à $D_{sa} = 1 - N_{sa}/2L$. Il existe d'autres mesures (comme l'apparentement) que nous verrons lors des analyses de données réelles. Il est certain que nous manquons de recul pour appréhender quelles mesures sont vraiment les meilleures et dans quelles situations.

Espèces haploïdes et loci liés au sexe

Certaines espèces sont haploïdes durant une certaine période (voire la totalité) de leur cycle de vie. Il se peut qu'elles soient étudiées (échantillonnées) durant cette phase. Bien entendu, il ne saurait être question d'étudier des hétérozygoties chez de telles espèces. Il n'en reste pas moins que des études de génétique des populations demeurent possibles, et en particulier le calcul de différenciation entre populations (F_{ST}). Selon le logiciel d'analyses étudié, il suffit juste de coder les données d'une manière spéciale (en général, on code les individus homozygotes à tous les loci).

Certaines espèces ont une reproduction sexuée avec des sexes séparés (espèces dites dioïques ou gonochoriques). Chez ces dernières, il peut arriver que le déterminisme du sexe soit chromosomique. Dans ce cas, l'un des deux sexes est déterminé par la possession de deux chromosomes identiques, alors que le second sexe est déterminé par une hétérogénéité à ce niveau, d'où le terme hétérogamétique qualifiant ce dernier. Dans le sexe hétérogamétique, un chromosome détermine le sexe (chromosome Y, ou chromosome W), il n'y aura en général que très peu de gènes et rarement les mêmes loci que sur l'autre chromosome. Dans certains cas, c'est même son absence qui détermine le sexe (mâles X0, par exemple). Il y aura donc, pour les individus hétérogamétiques, haploïdie de fait pour les loci situés sur les chromosomes sexuels (en général donc sur l'X ou le Z). Les mammifères et les drosophiles (Chippindale et Rice, 2001) ont par exemple un déterminisme du sexe XY (femelles XX et mâles XY). C'est également le cas de la plupart des tiques Argasidae et *Ixodes* (Kiszewski *et al.*, 2001). Les oiseaux ont, quant à eux, un déterminisme du type ZW/ZZ (femelles ZW). C'est également ainsi que le sexe est déterminé chez les schistosomes (Hirai et LoVerde, 1995). Chez certaines espèces de nématodes (Šnabel *et al.*, 2000), chez la plupart des tiques des genres *Dermacentor*, *Amblyomma* et *Rhipicephalus* et l'espèce *Ixodes holocyclus* (Kiszewski *et al.*, 2001) ainsi que chez certains pucerons (Caillaud *et al.*, 2002) le système est du type XX (femelles) et X0 (mâles). Dans ces cas, ces loci sont tout de même utilisables en ce qui concerne les études de différenciation ou de diversité génique (selon le logiciel d'analyse, on les code homozygotes) sauf pour l'estimation de F_{IS}, pour laquelle les données à ces loci doivent bien évidemment être éliminées (codées en données manquantes). C'est ce qui a été fait pour la tique *Ixodes ricinus* (De Meeûs *et al.*, 2002a) ou pour les

mouches tsé-tsé (Camara *et al.*, 2006 ; Ravel *et al.*, 2007). Il est cependant clair que ce n'est pas idéal et l'utilisation de loci autosomaux devrait idéalement être favorisée.

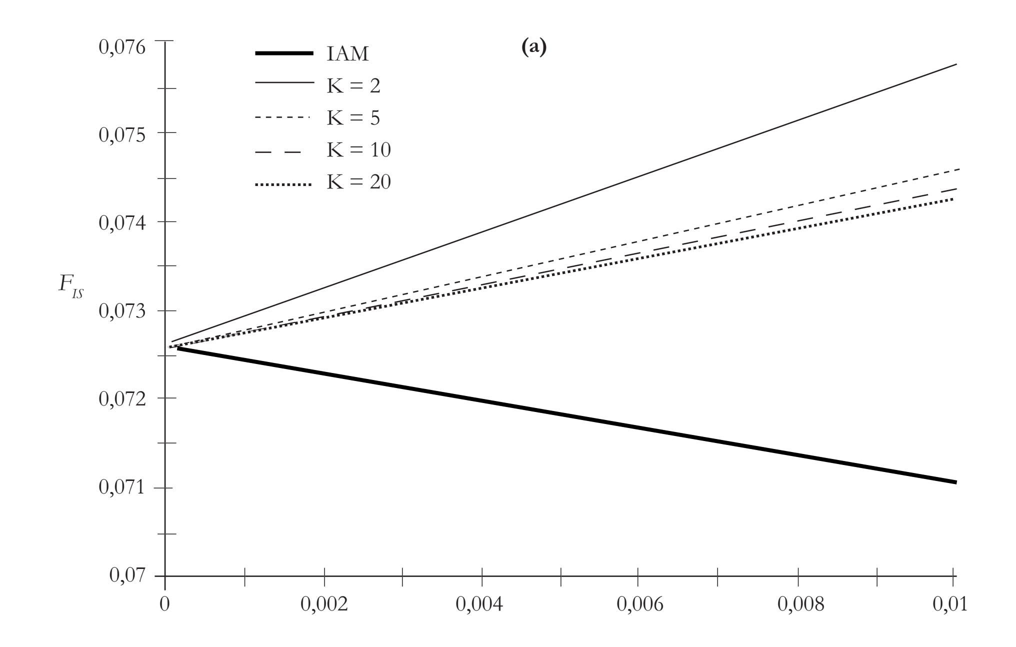
Le problème de l'homoplasie

Comme nous l'avons déjà vu, les marqueurs génétiques polymorphes dont nous avons besoin pour analyser nos populations naturelles correspondent rarement à des loci à nombre infini d'allèles. C'est par exemple le cas des allozymes pour lesquels un grand nombre de mutations différentes sont confondues dans un seul allèle. C'est aussi vrai pour les microsatellites les plus polymorphes, de par les contraintes issues du mécanisme mutationnel de ces séquences particulières d'ADN, beaucoup d'allèles sont identiques par état sans être identiques par descendance (ou ascendance en fonction de la direction vers laquelle nous regardons). On parle alors d'homoplasie. Pour certains, ce phénomène est rédhibitoire en génétique des populations. Tout d'abord, en ce qui concerne le F_{IS}, il a été démontré que ce dernier est virtuellement indépendant du processus de mutation (Rousset, 1996). En ce qui concerne les mesures de différenciation, nous avons vu avec l'équation (25) que le biais du F_{ST} est proportionnel à $K/(K - 1)$ quand K est le nombre d'allèles possibles. Ce biais est donc faible pour des nombres raisonnables d'allèles. La figure 9 illustre bien la modestie de l'influence de l'homoplasie sur les paramètres courants utilisés en génétique des populations.

Cette influence, quasi nulle sur le F_{IS} (notez la faiblesse de l'échelle), devient rapidement négligeable dès que le nombre d'allèles possibles dépasse 5, voire même 2 quand les taux de mutation sont au-dessous de 10^{-4}. Si on ajoute que les variances des estimateurs de ces paramètres sont telles qu'il n'est pas raisonnable d'espérer une précision en deçà de deux décimales, l'homoplasie n'est absolument pas un problème pour le F_{IS} et donc pour les inférences liées au système de reproduction, et ne représente qu'un problème modeste pour le F_{ST} dans les cas à deux allèles (qu'il vaut donc mieux éviter), de toutes manières assez rares, surtout chez les marqueurs microsatellites. Donc, si les loci homoplasiques sont bien évidemment à éviter pour toute étude phylogénétique ou assimilée, il n'y a aucune raison valable de les écarter pour des études de génétique des populations.

Structuration à plus de trois niveaux

La situation classique à trois niveaux, individus, sous-population, population totale ne correspond bien évidemment pas à tous les cas de figure. Il peut, par exemple exister plus de niveaux. Si j'échantillonne plusieurs parasites par individu hôte, avec plusieurs hôtes dans plusieurs habitations de plusieurs villages, on voit bien que l'on peut avoir autant de niveaux pertinents de structuration potentiels. On peut alors subdiviser l'échantillon en autant de sous-échantillons qu'il est nécessaire en supprimant l'influence des niveaux potentiellement confondants. Par exemple, l'effet individu hôte



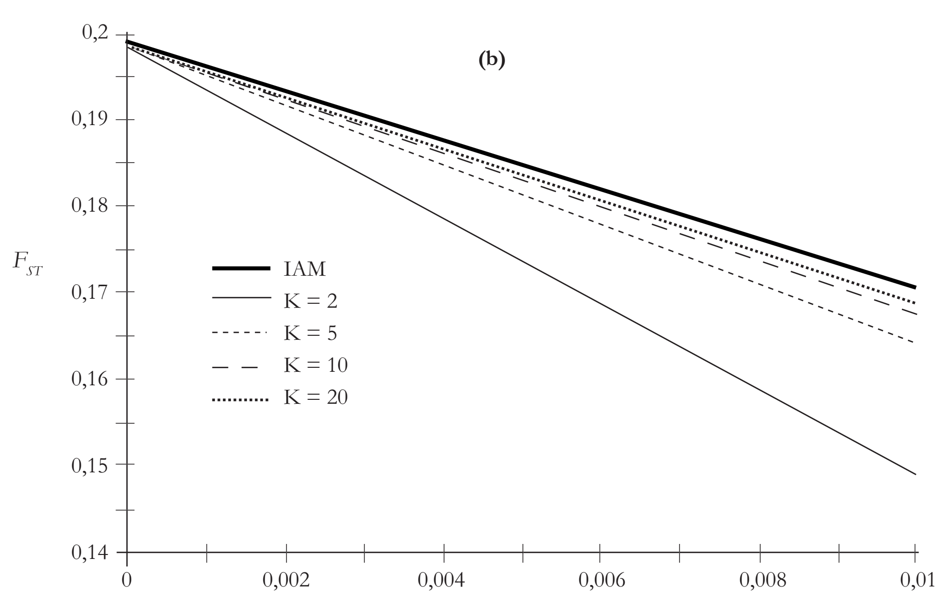


Figure 9
Illustration de l'influence du nombre d'allèles possibles (*K*) sur les valeurs attendues des *F* statistiques de Wright, pour différents taux de mutation (*u*). Les valeurs sont obtenues pour un modèle en îles infini avec *N* = 20 individus par sous-population, un taux d'autofécondation de *s* = 0,2 et un taux de migration de *m* = 0,05 (IAM = nombre infini d'allèles).

peut être étudié en considérant chaque habitation séparément et en calculant un F_{ST}, alors que l'analyse de l'effet habitation se fera dans chaque village (séparé) en ne gardant les parasites que d'un seul individu hôte par habitation, en calculant de nouveau un F_{ST} et en regardant de combien celui-ci est différent du précédent. Cette tâche devient rapidement irréalisable et de toutes façons assez contestable, car il existe une solution beaucoup plus élégante. Le logiciel HierFstat (Goudet, 2005, disponible à http://www2.unil.ch/popgen/softwares/hierfstat.htm) offre la possibilité d'estimer les F hiérarchiques pour toute structure hiérarchisée de population et ce en une seule analyse, comme cela a été utilisé avec profit dans Trouvé *et al.* (2005) ou Nébavi *et al.* (2006). Dans notre exemple, on aurait donc obtenu $F_{\text{Parasites_Hôte}}$, $F_{\text{Hôte_Habitation}}$, $F_{\text{Habitation_Village}}$, $F_{\text{Village_Total}}$. Ces différents F sont calculés et estimés suivant des principes analogues à ceux décrits en (21) et (40) et donnent donc les parts respectives des différents niveaux hiérarchiques dans la partition de la variation génétique. Par exemple, $F_{\text{Habitation_Village}}$ donne la différenciation génétique entre habitations dans chaque village en corrigeant pour l'effet individu hôte.

Ceci est plus important qu'il n'y paraît. S'il y a un effet significatif de l'individu hôte et que les parasites de ces derniers sont confondus, le calcul d'un F_{ST} entre habitations sera nécessairement biaisé, car l'effet individu hôte y sera nécessairement inclus (les habitants de différentes habitations sont différents).

Probabilités (ou indices) d'assignement

Le génotype multilocus d'un individu peut permettre de calculer la probabilité avec laquelle cet individu appartient à une sous-population donnée (Rannala et Mountain, 1997 ; Waser et Strobeck, 1998 ; Cornuet *et al.*, 1999). Cette probabilité correspond alors simplement à la probabilité multinomiale attendue d'observer un génotype particulier compte tenu des fréquences des allèles dans la sous-population. La précision de cette probabilité dépend nécessairement de celle avec laquelle les fréquences alléliques sont estimées et donc de la taille de l'échantillon. Idéalement, la mesure devrait être effectuée à partir d'au moins 30 individus génotypés sur au moins 10 loci variables dans chaque sous-population. Cette probabilité est conventionnellement appelée indice d'assignement. Un individu présentant une faible valeur correspondra probablement à un immigrant récent. En comparant l'indice d'un individu pour différentes populations, on peut aussi essayer de détecter sa population d'origine, pour autant que cette population ait été échantillonnée bien entendu.

On peut aussi analyser ces indices pour détecter des individus parasites mieux adaptés à telle ou telle autre espèce d'hôte (races d'hôte) en comparant les indices d'assignement des individus parasites aux différentes espèces hôtes sur lesquelles ils ont été prélevés ainsi qu'aux différentes localités de prélèvements où les échantillons ont été effectués (voir par exemple McCoy *et al.*, 2005). On utilise également ces indices pour étudier des catégories d'individus (femelles versus mâles, hôtes parasités versus sains).

L'indice d'assignement (AI) (Paetkau *et al.*, 1995) d'un individu k échantillonné dans une population l correspond à la probabilité que ce génotype soit retrouvé par chance dans cette population l, compte tenu des fréquences alléliques dans cette population (fréquences calculées en y incluant l'individu focal). Pour un locus donné, si les fréquences des allèles i et j dans la population l sont p_i et p_j respectivement, la probabilité d'appartenir à cette population est de p_{il}^2 pour les homozygotes et $2p_{il}p_{jl}$ pour les hétérozygotes. Les probabilités obtenues pour chaque locus (p_l) doivent être multipliées entre elles pour obtenir $AI = \prod_{l=1}^{l=L} p_l$ sur l'ensemble des L loci.

D'autres méthodes permettent de calculer un indice d'assignement. Par exemple, Piry *et al.* (2004) utilisent une méthode bayésienne, avec exclusion de l'individu focal.

Par ailleurs, on peut ne pas souhaiter comparer des populations, mais plutôt des catégories d'individus dans les populations. On peut se demander par exemple si les hôtes parasités sont plus ou moins bien assignés que les sains, ce qui refléterait une modification du comportement des individus sous l'effet du parasitisme. On peut désirer savoir si les individus mâles ont le même comportement que les femelles. Il est alors intéressant de s'affranchir du biais imposé par le degré de polymorphisme contenu dans chaque sous-population. On utilise alors une version centrée de AI (AI_c) en retranchant de AI la moyenne de la population, après transformation Log (pour minimiser l'impact des trop petites valeurs) (Favre *et al.*, 1997). Il en résulte que l'AI_c moyen de chaque sous-échantillon est centré sur 0 et que les individus avec un AI_c négatif correspondent à des individus moins bien assignés à leur population d'origine que la moyenne des individus échantillonnés sur place. Cette dernière méthode est celle que l'on trouve dans Fstat 2.9.3. (Goudet, 2002) alors que celle de Piry *et al.* peut être trouvée dans GeneClass 2 (Piry et Alapetite, 2003) téléchargeable à http://www.montpellier.inra.fr/URLB/.

Pour plus de détails sur les indices d'assignement, on pourra consulter Manel *et al.* (2005).

LES DÉSÉQUILIBRES DE LIAISON

Comme mentionné précédemment, il est indispensable de travailler à partir de l'information recueillie sur plusieurs loci. Un problème supplémentaire peut alors survenir sous la forme d'une corrélation entre les allèles de différents loci. Nous allons maintenant nous intéresser au polymorphisme à deux loci. Si ces deux loci polymorphes (au moins deux allèles chacun) sont indépendants dans une population qui suit les hypothèses de Hardy-Weinberg, on s'attend, à l'équilibre, à retrouver une association aléatoire entre les allèles des deux loci. Comme en général on n'a pas

accès à la phase des gamètes, on regarde cela au niveau des génotypes des individus diploïdes. Donc si D_1, H_1 et R_1, et D_2, H_2 et R_2 sont les fréquences des génotypes 11, 12 et 22 aux loci 1 et 2 respectivement (on supposera pour simplifier qu'il n'y a que deux allèles et donc trois génotypes par locus), on s'attend alors à trouver des fréquences de génotypes aux deux loci suivants :

$$\begin{cases} f(11_11) = D_1D_2\ ;\ f(11_12) = D_1H_2\ ;\ f(11_22) = D_1R_2\ ; \\ f(12_11) = H_1D_2\ ;\ f(12_12) = H_1H_2\ ;\ f(12_22) = H_1R_2\ ; \\ f(22_11) = R_1D_2\ ;\ f(22_12) = R_1H_2\ ;\ f(22_22) = R_1R_2\ ; \end{cases} \quad (43)$$

Si les fréquences bilocus observées diffèrent de celles décrites dans le système d'équations (43), on dit qu'il y a déséquilibre de liaison. Il s'agit d'un déséquilibre statistique uniquement, car rien ne prouve que les loci sont liés physiquement (proches sur le même chromosome). La liaison physique peut bien sûr représenter une cause possible d'un déséquilibre statistique de liaison entre deux loci, mais d'autres phénomènes peuvent conduire à une telle observation. Les systèmes de reproduction fermés (autofécondation ou mieux clonalité) sont par eux-mêmes susceptibles de générer d'importants déséquilibres de liaison entre tous les loci du génome. La sélection naturelle, quand elle favorise des combinaisons spécifiques d'allèles à différents loci, ou quand elle est épistatique (Ohta, 1982 ; Chippindale et Rice, 2001), peut elle aussi être rendue responsable de la liaison entre certains loci. Enfin, et ce n'est pas la moins importante des causes à signaler, l'interaction entre mutation, dérive et migration peut elle aussi générer des déséquilibres de liaison entre loci, en particulier dans les populations très structurées (petites sous-populations échangeant peu de migrants). Comme nous n'avons généralement pas accès à la phase haploïde (gamétique) des organismes étudiés (sauf chez des haploïdes évidemment), l'estimation du déséquilibre de liaison ne peut se faire que de façon composite (12_12 ne peut être distingué de 21_12) (Weir, 1979, 1996). Parce que les systèmes de reproduction sexués fermés, comme l'autofécondation ou la parthénogénèse, ou encore les systèmes clonaux (reproduction végétative) conduisent à un déséquilibre global, certains auteurs ont développé des mesures multilocus du déséquilibre de liaison (par exemple, Agapow et Burt, 2001). Il est cependant important de noter ici que le comportement de ces différentes mesures dans différentes conditions de populations n'a été que peu étudié jusqu'à présent malgré l'importance soulignée de telles études (De Meeûs et Balloux, 2004). L'absence de déséquilibre de liaison est une hypothèse souvent mise en avant, car certaines analyses statistiques considèrent l'information apportée par les différents loci comme indépendante. Un déséquilibre de liaison fort risquant d'apporter une redondance conduisant à un risque d'erreur de décision (traité dans le chapitre suivant), il est souvent plus « confortable » de pouvoir écarter ce problème, tout en sachant qu'il ne peut exister de population exempte de déséquilibre de liaison.

3 Tests statistiques

BASES

Intuitivement, on sait qu'un échantillon ne sera jamais une représentation absolument fidèle de ce qui existe dans la population entière. Il en découle que l'échantillonnage provoque une déviation des estimateurs que l'on peut calculer (variance d'échantillonnage). On aura donc rarement, même dans une population échantillonnée strictement conforme à Hardy-Weinberg, un F_{IS} estimé exactement égal à 0, pareil pour le F_{ST}, pour les déséquilibres de liaison, l'isolement par la distance ou n'importe quel autre paramètre. Le test statistique est là pour nous aider à prendre une décision quant à la disparité observée entre les données et l'attendu. La différence observée peut-elle être expliquée par le hasard et avec quelle probabilité ? Le but d'un test statistique sera donc de fournir une réponse à cette question, en donnant un critère, la valeur *P* ou *P*-value[4] du test, ou risque de première espèce ou encore probabilité de se tromper en répondant par la négative à cette question (appelée hypothèse nulle). On peut ajouter ici qu'en génétique des populations, la variance d'échantillonnage est d'autant plus importante à prendre en compte que l'échantillonneur lui-même n'arrive qu'à la fin d'un processus d'échantillonnage qui a lieu naturellement : échantillonnage parmi les gamètes disponibles pour fabriquer les zygotes ; échantillonnage des zygotes qui participeront à la reproduction suivante.

L'hypothèse nulle

Comme son nom l'indique, c'est une hypothèse qui stipule qu'il ne se passe rien, ou que la population est conforme à une norme, un modèle préétabli (ou modèle nul), par exemple les fréquences génotypiques sont conformes à Hardy-Weinberg, ou les deux populations ont les mêmes fréquences alléliques, ou encore le F_{IS} (ou le F_{ST}) n'est pas différent de 0. On nomme cette hypothèse sous le diminutif H_0. L'hypothèse alternative, ou H_1, peut être indéfinie (par exemple, le F_{IS} est différent de 0) ou au contraire définie (ou orientée) (exemple, le F_{IS} est plus grand que 0). Dans ce dernier cas, on parle de test unilatéral qui, comme nous le verrons, est en général plus puissant que le premier (ou test bilatéral), sauf si on se trompe de direction (voir plus loin).

[4] J'utiliserai cet anglicisme tout au long de ce manuel, car il est devenu d'usage courant, comme week-end, mail ou web.

Qu'est-ce qu'un test statistique ?

Un test statistique consiste en un calcul plus ou moins compliqué de la probabilité avec laquelle le hasard (et seulement lui) nous permet d'expliquer la déviation observée dans un échantillon par rapport à ce qui est attendu sous H_0.

Prenons un exemple très simple. Je tire deux fois à pile ou face. Je peux soit obtenir deux piles avec la probabilité $(½)^2$, soit un pile et une face avec la probabilité ½ et deux faces avec la probabilité $(½)^2$. Je joue et obtiens deux faces. Nous allons procéder à trois tests statistiques.

– Test unilatéral 1 :

H_0 : la pièce est bien équilibrée, ce que j'ai observé n'est pas significativement différent de l'attendu ½ / ½.

H_1 : la pièce n'est pas bien équilibrée, j'obtiens plus de faces qu'attendu.

– Test unilatéral 2 :

H_0 : la pièce est bien équilibrée, ce que j'ai observé n'est pas significativement différent de l'attendu ½ / ½.

H_1 : la pièce n'est pas bien équilibrée, j'obtiens moins de faces qu'attendu.

– Test bilatéral :

H_0 : la pièce est bien équilibrée, ce que j'ai observé n'est pas significativement différent de l'attendu ½ / ½.

H_1 : la pièce n'est pas bien équilibrée et j'obtiens un résultat significativement différent de l'attendu.

– Pour le test unilatéral 1, la probabilité d'obtenir par hasard autant ou plus de faces est égale à P_{u1} = (Somme des probabilités d'obtenir autant ou plus que deux faces)/(Somme des probabilités totales obtenues) = $(½)^2/1$. Donc la P-value du test est $P_{u1} = 0{,}25$.

– Pour le test unilatéral 2, la probabilité d'obtenir par hasard plus ou autant de piles est égale à P_{u2} = (Probabilité de zéro pile + Probabilité de un pile et une face + Probabilité de deux piles)/(Somme totale) = $((1/2)^2 + 1/2 + (1/2)^2)/1$. Donc $P_{u2} = 1$.

– Pour le test bilatéral P_b = (probabilité d'avoir deux faces ou deux piles) = $((1/2)^2 + (1/2)^2)/1 = 0{,}5$.

Plusieurs choses peuvent ici être signalées. Tout d'abord, la plus basse des probabilités obtenues est 0,25. Ce qui illustre la faible puissance du test due à la faiblesse de l'échantillon. Il est difficile de prouver quelque chose avec de trop petits échantillons. Ensuite, on voit bien qu'on est beaucoup plus puissant en unilatéral si on teste dans la bonne direction, et beaucoup moins quand on teste dans la mauvaise. Il faut décider du test que l'on fait, bilatéral ou unilatéral et dans quelle direction,

avant de faire le test. Il faut donc bien se poser la question avant, pas après. Si aucune information ne permet de définir dans quelle direction le signal doit avoir lieu, il faut systématiquement procéder à un test bilatéral. Par contre, si on est certain de la direction que le signal est censé prendre, alors le test unilatéral s'impose.

Par exemple, je mesure le F_{IS} à partir d'un échantillon quelconque et j'obtiens une valeur légèrement plus grande que 0. Je pose tout d'abord mon hypothèse nulle :

H_0 : F_{IS} n'est pas significativement différent de 0. Le test statistique va donc consister à calculer, compte tenu du nombre de loci sur lequel la mesure a été faite, le degré de polymorphisme de ces différents loci (nombre d'allèles, leur distribution) et le nombre d'individus génotypés, la probabilité d'avoir obtenu un F_{IS} aussi extrême[5] ou plus extrême que celui observé, sous l'hypothèse d'une rencontre au hasard des gamètes dans la population d'où ont été tirés les individus génotypés (panmixie). Le test par défaut est unilatéral et le plus souvent pour les valeurs positives (H_1 : $F_{IS} > 0$), car la plupart des facteurs influençant ce paramètre génèrent des déficits en hétérozygotes (autofécondation, effet Wahlund…). Cependant, dans certains cas, comme celui d'une reproduction clonale partielle, on s'attend à des déviations dans les deux directions (Balloux *et al.*, 2003 ; De Meeûs *et al.*, 2006). Dans ce cas, et comme les logiciels disponibles ne donnent pas de tests bilatéraux, il faut cumuler les résultats des deux tests unilatéraux ($F_{IS} > 0$ et $F_{IS} < 0$) en sommant $P_{\min} + 1 - P_{\max}$, où $P_{\min}$ et $P_{\max}$ correspondent à la plus petite (test unilatéral le plus puissant) et la plus grande des deux *P*-values des deux tests ou, si $P_{\max} = 1$ comme c'est le cas ici, en doublant $P_{\min}$.

Risques de première et de seconde espèce

En règle générale, on considère (arbitrairement) qu'un test est significatif quand la *P*-value à laquelle il est associé est inférieure ou égale à 0,05. Mais dans certains cas (que nous verrons plus loin), il peut s'avérer nécessaire d'être plus sévère et de baisser ce seuil. Personnellement, je me sens plus à l'aise avec une *P*-value < 0,01 pour rejeter H_0 et une *P*-value > 0,1 pour l'accepter. Le seuil à partir duquel on décide qu'une statistique est significative (rejet de H_0) est appelé risque de première espèce ou erreur de type I et noté α. Il s'agit du risque de se tromper en rejetant H_0 quand elle est vraie. Le risque de seconde espèce, ou erreur de type II, noté β, correspond au risque de se tromper en acceptant l'hypothèse nulle quand elle est fausse. Ce risque, qui est fonction de la puissance du test, est très rarement connu mais peut être appréhendé dans certaines circonstances. L'exemple du pile ou face ci-dessus est typiquement un cas où β est nécessairement très grand puisque, même si la pièce est truquée, on ne pourra jamais le détecter en ne faisant que deux essais.

[5] Les valeurs du F_{IS} peuvent s'écarter de 0 en se montrant fortement négatives ou fortement positives.

LE PRINCIPE DES RANDOMISATIONS

Dans la plupart des situations rencontrées en génétique des populations naturelles (si ce n'est toutes), il ne sera pas possible de procéder au calcul des probabilités exactes telles que dans l'exemple du pile ou face. Cependant, l'utilisation de programmes informatiques va nous permettre, sans beaucoup d'effort, d'estimer avec une excellente approximation, ces *P*-values. Il s'agit de procédures de ré-échantillonnage ou randomisations. Ces procédures se regroupent en deux grands types. Celles du premier type visent à obtenir un intervalle de confiance de l'estimateur étudié (par exemple, le F_{IS}), l'autre vise à simuler des populations suivant l'hypothèse nulle afin de pouvoir comparer la valeur observée à celles qu'on peut attendre sous H_0 (obtenues par simulation).

La plupart des tests décrits dans ce manuel sont disponibles dans le logiciel Fstat 2.9.3. (Goudet, 2002, mise à jour de Goudet, 1995), qui est très convivial. D'autres logiciels sont aussi utiles :

– Genepop 3.4. (Raymond et Rousset, 2003, mise à jour de Raymond et Rousset, 1995b), Genepop 4 (Rousset, 2008), moins convivial, mais qui est le seul à proposer certaines procédures très utiles (comme celles testant des isolements par la distance entre individus) et leur version web ;

– Genetix 4.03, très convivial, en français qui propose des AFC (analyses factorielles des correspondances) ;

– MSA (Dieringer et Schlötterer, 2002), pas très convivial, mais qui propose différents calculs de distances génétiques.

Il en existe bien sûr bien d'autres que nous utiliserons dans la 2e partie de ce manuel « Applications à des exemples concrets », mais avec ces trois-ci on peut déjà faire énormément de choses. Ajoutons que ces logiciels sont téléchargeables gratuitement (voir en annexe les liens), chose à ajouter au crédit de leurs auteurs. Nous reviendrons sur d'autres logiciels au moment où nous en aurons besoin.

Mais avant tout, il y a Create (Coombs *et al.*, 2008) qui permet, à partir d'un fichier texte ou Excel avec toutes les données brutes, de convertir ces données dans un format adéquat pour la plupart des logiciels de génétique des populations. Au moment où je corrige mon manuscrit, Tatiana Giraud m'apprend qu'il en existe un autre PGD-Spider (Lischer et Excoffier, 2012), apparemment assez convivial, mais que je n'ai encore jamais utilisé.

Intervalles de confiance de bootstrap et jackknife

Le bootstrap

Il s'agit d'un rééchantillonnage répété avec remise. On sélectionne au hasard un des réplicats et, après avoir noté sa valeur, on le remet et ainsi de suite jusqu'à obtention

d'autant de mesures qu'il y a de réplicats dans l'échantillon. La procédure est répétée un grand nombre de fois (5 000 pour Fstat). On obtient ainsi une distribution de 5 000 valeurs possibles. En excluant les 2,5 % (0,5 %) plus petites et les 2,5 % (0,5 %) plus grandes de ces valeurs, on obtient l'intervalle de confiance à 95 % (99 %). Ces notions seront plus claires avec la description du bootstrap sur les loci et sur les populations.

Bootstrap sur les loci

On rééchantillonne au hasard et avec remise les *k* différents loci disponibles, jusqu'à en avoir *k* sur lesquels on recalcule l'estimation du paramètre (F_{IS} ou F_{ST}). On recommence l'opération un très grand nombre de fois (5 000 fois). Notons que, puisqu'il s'agit d'un échantillonnage des loci avec remise, on peut obtenir plusieurs fois le même locus. On obtient ainsi une distribution des valeurs obtenues sur les 5 000 bootstraps. Il ne reste plus ensuite qu'à regarder les valeurs obtenues de part et d'autre de cette distribution pour obtenir un intervalle de confiance. Par exemple pour 5 000 bootstraps, la valeur obtenue avant les 2,5 % les plus fortes et après les 2,5 % les plus faibles nous donne l'intervalle de confiance à 95 % (voir la figure 10).

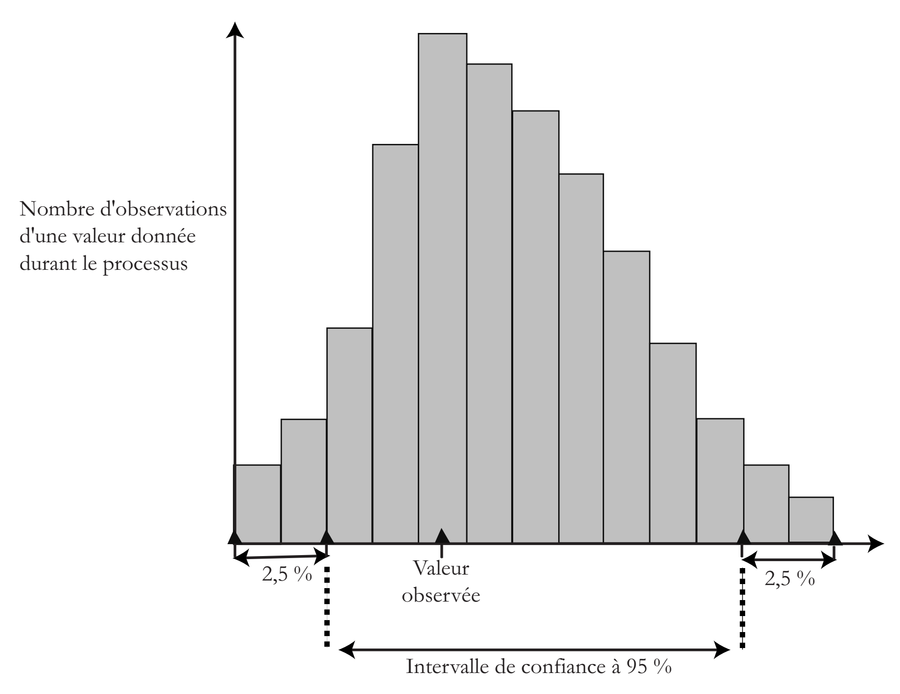


Figure 10
Représentation graphique de l'obtention de l'intervalle de confiance à 95 % d'une mesure à partir de la technique du bootstrap.

Dans la figure 10, on voit que la valeur observée n'est pas centrée, car le bootstrap génère des distributions décalées. Cette procédure sert à comparer des statistiques F entre différents échantillons ou groupes. Elle n'est pas très puissante, mais fournit la possibilité de faire des graphiques élégants. En général, on l'accompagne d'un autre test plus puissant, par exemple un test de Wilcoxon pour données appariées (par loci si les loci étudiés sont les mêmes) ou un test de Kruskal-Wallis si les loci ne sont pas les mêmes (les deux tests sont implémentés dans tous les logiciels de statistiques). Attention, si les loci ne sont pas les mêmes, la différence observée entre groupes pourra provenir des loci et non des groupes. D'une manière générale, il vaut mieux s'assurer de travailler avec les mêmes loci et que ces derniers soient en nombre suffisant (au moins sept). De toutes les façons, un bootstrap sur les loci ne commence à avoir du sens qu'à partir de quatre loci, et est vraiment puissant bien au-delà (voir Raymond et Rousset, 1995a pour discussion).

Bootstrap sur les populations

C'est exactement le même principe que le précédent sauf que ce sont les populations (ou ce que l'on considère comme telles, les sous-échantillons) qui sont ici rééchantillonnées. Attention, on ne peut pas faire cela pour le F_{ST}[6]. Seul donc le F_{IS} est concerné. Cette procédure permet de comparer les loci entre eux. Il est en effet important de vérifier si les différents loci convergent vers le même signal, car sinon il sera utile de déterminer les causes responsables des discordances entre loci.

Le jackknife

Jackknife sur les loci

Ici, il s'agit de prendre chaque locus un à un et de calculer la valeur du F sur ceux qui restent. On obtient ainsi k valeurs sur lesquelles on peut calculer une moyenne et une variance et donc une erreur standard. L'erreur standard d'une statistique x évaluée sur n mesures se calcule en fonction de sa variance $s^2(x)$ [voir équation (28)] et de n selon la formule :

$$s_{\bar{x}} = \sqrt{\frac{s^2(x)}{n}} \tag{44}$$

Il est ensuite facile de calculer à partir de là l'intervalle de confiance voulu (IC), en faisant l'hypothèse que la distribution des jackknives suit une distribution normale (ce qui n'est probablement pas tout à fait exact, mais passons).

$$IC = F \pm t_{n-1,\alpha} s_{\bar{x}} \tag{45}$$

où F est le F de Wright étudié, et $t_{n-1,\alpha}$ le paramètre de la loi normale pour n-1 degré de liberté (n correspond ici au nombre de loci) et au seuil α (α = 0,05 pour un IC

[6] Le bootstrap rééchantillonne avec remise et peut donc dans ce cas rééchantillonner plusieurs fois le même sous-échantillon. Mesurer la différenciation entre ces échantillonnages strictement identiques n'a aucun sens.

de 95 %). Cette valeur du t peut être retrouvée à partir d'une table (tabl. 3) ou d'un programme informatique.

Pour n très grand et $\alpha = 0,05$, on a classiquement $t = 1,96$. Sous Excel, la commande est = LOI.STUDENT.INVERSE(A1;A2-1) où A1 correspond à la case de coordonnées de la colonne A, 1re ligne où vous avez tapé la valeur pour α et A2 est la case où vous avez tapé la valeur du nombre de réplicas, d'où on retranche 1 pour avoir le degré de liberté.

Jackknife sur populations

Même chose que pour les loci, mais avec les sous-échantillons. Notons que le F_{ST} peut se prêter à cette procédure ici, puisqu'on a toujours n-1 sous-échantillons sur lesquels calculer un F_{ST}, ce qui n'est pas garanti par le bootstrap (le même sous-échantillon peut être échantillonné n fois par la procédure).

Applications numériques pour le jackknife

Supposons un jeu de données sur huit échantillons avec cinq loci. Sur l'ensemble des loci $F_{ST} = 0,004$, et pour le Locus 1 $F_{ST1} = 0,002$. Le jackknife sur loci (cinq valeurs) donne une erreur standard de StdErrLoci(F_{ST}) = 0,003. Le jackknife sur populations (huit valeurs) donne une erreur standard de StdErrPop(F_{ST1}) = 0,001 pour le Locus 1. Alors, les intervalles de confiance à 95 % de F_{ST} et F_{ST1} seront :

$$\begin{cases} \mathrm{CI}(F_{ST}) = F_{ST} \pm t_{0.05,\gamma L}\mathrm{StdErrLoci}(F_{ST}) \\ \mathrm{CI}(F_{ST1}) = F_{ST1} \pm t_{0.05,\gamma P}\mathrm{StdErrLoci}(F_{ST1}) \end{cases}$$

avec $\gamma L = 5 - 1 = 4$ et $\gamma P = 8 - 1 = 7$ correspondant aux degrés de liberté des procédures de jackknife sur loci et populations respectivement. En utilisant le tableau 3, nous obtenons alors $t_{0,05,\gamma L} = 2,776$ et $t_{0,05,\gamma P} = 2,365$, d'où l'on tire que CI($F_{ST}$) = 0,004 ± 0,008 et CI($F_{ST1}$) = 0,002 ± 0,002.

En règle générale, je préfère utiliser le bootstrap car il ne fait pas d'hypothèse, mais si je souhaite obtenir des intervalles de confiance du F_{ST} pour les différents loci, je suis bien obligé de le faire avec un jackknife sur populations.

Comme pour le bootstrap, il faut au minimum cinq réplicats (loci ou populations) pour qu'un jackknife soit effectué par Fstat.

Mise en garde

Comme nous l'avons vu, le nombre de réplicats à rééchantillonner doit respecter une valeur minimum. Il est nécessaire qu'il y ait au moins cinq loci et/ou sous-échantillons pour que ces procédures soient effectuées. Il est également nécessaire que ces réplicats soient suffisamment variables, un locus monomorphe ou presque pas variable ne pourra pas offrir un réplicat digne de ce nom, même si Fstat effectue la procédure sans problème (c'est-à-dire sans vous prévenir qu'il y a potentiellement un souci).

Tableau 3
Valeurs du *t* pour différents degrés de liberté (*n*-1) au seuil α = 0,05.

n-1	*t*(α = 0,05)	*n*-1	*t*(α = 0,05)	*n*-1	*t*(α = 0,05)
1	12,706	21	2,080	45	2,014
2	4,303	22	2,074	50	2,009
3	3,182	23	2,069	55	2,004
4	2,776	24	2,064	60	2,000
5	2,571	25	2,060	65	1,997
6	2,447	26	2,056	70	1,994
7	2,365	27	2,052	80	1,990
8	2,306	28	2,048	90	1,987
9	2,262	29	2,045	100	1,984
10	2,228	30	2,042	110	1,982
11	2,201	31	2,040	120	1,980
12	2,179	32	2,037	130	1,978
13	2,160	33	2,035	140	1,977
14	2,145	34	2,032	150	1,976
15	2,131	35	2,030	200	1,972
16	2,120	36	2,028	250	1,970
17	2,110	37	2,026	300	1,968
18	2,101	38	2,024	400	1,966
19	2,093	39	2,023	500	1,965
20	2,086	40	2,021	1000	1,962

Les permutations

Il s'agit ici de simuler l'hypothèse nulle un grand nombre de fois avec les données. Le programme informatique va utiliser les données (c'est-à-dire les allèles ou les individus des différents sous-échantillons) pour simuler H_0, mesurer la valeur obtenue sous H_0, recommencer un très grand nombre de fois afin d'obtenir une distribution des valeurs possibles sous H_0. La *P*-value du test correspond donc simplement à la proportion des cas où une valeur aussi grande ou plus grande (unilatéral 1), aussi petite ou plus petite (unilatéral 2), aussi extrême (bilatéral) que la valeur observée a été obtenue dans cette distribution.

Il existe deux grands types de randomisations : les permutations et les chaînes de Markhov. Les deux sont basées sur le principe de Monte Carlo. Le nom provient d'un clin d'œil de Metropolis à son collègue Stan Ulam et n'est pas sans rapport avec les jeux de hasard ayant cours dans la capitale de Monaco (voir METROPOLIS, 1987)[7]. Il s'agit de rééchantillonnages aléatoires (ou pseudo-aléatoires) des données.

La procédure de permutation correspond à la méthode utilisée dans Fstat (GOUDET, 1995). Il s'agit en fait de simuler l'hypothèse nulle un grand nombre de fois (par exemple, 10 000 fois) à partir des données existantes (l'échantillon). Par exemple, dans chaque sous-échantillon et pour chaque locus, les allèles de tous les individus sont réassociés deux à deux au hasard pour reformer des individus fictifs obtenus selon l'hypothèse de panmixie locale. Un F_{IS}, obtenu donc par hasard sous H_0, est calculé et on recommence la même chose un très grand nombre de fois. La proportion de fois qu'un F_{IS} aussi grand ou plus grand que celui observé dans les données est apparu au cours du processus correspond à la *P*-value du test (H_1 étant ici $F_{IS} > 0$). Selon un principe analogue, la différenciation (H_0 : les individus se distribuent de façon aléatoire dans les différents sous-échantillons) est testée en assignant chaque individu aléatoirement dans les différents sous-échantillons, un F_{ST} obtenu sous H_0 est calculé et le processus répété. La proportion de fois qu'un F_{ST} (ou une autre statistique reflétant la distribution des fréquences alléliques entre sous-échantillons, comme nous le verrons plus loin) simulé sous H_0 a été aussi grand ou plus grand que l'observé procure la *P*-value du test.

La chaîne de Markhov correspond à la procédure utilisée dans Genepop (RAYMOND et ROUSSET, 1995b). Le principe en est le suivant. Il s'agit de définir une trajectoire aléatoire entre différents tableaux de contingences possibles et dont les sommes marginales sont identiques au tableau des données réelles. La probabilité d'apparition (sous H_0) de chacun des tableaux ainsi obtenus est comparée à celle correspondant au tableau de contingence observé. La probabilité du test est obtenue en comptant le nombre de fois qu'une probabilité s'est montrée inférieure ou égale à celle du tableau observé et en divisant cette valeur par le nombre total de tableaux générés durant le processus. Une description mieux détaillée est présentée dans ROUSSET et RAYMOND (1997).

Les *P*-values obtenues par ces méthodes constituent d'excellentes approximations des probabilités exactes, pour peu que l'on ait pris garde de mettre en œuvre un nombre suffisant de ces randomisations (un minimum de 1 000 à 10 000 pour les permutations et 10^6 à 10^7 pour les chaînes de Markhov), ce qui ne devrait pas représenter un problème avec les micro-ordinateurs d'aujourd'hui.

[7] *I suggested an obvious name for the statistical method-a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte-Carlo." The name seems to have endured.*

TESTER LA PANMIXIE LOCALE

Tester le F_{IS}

La panmixie locale peut se tester en prenant les allèles présents dans chaque sous-échantillon et en les réassociant au hasard à l'intérieur de ces sous-populations et ce dans toutes les sous-populations. On mesure alors le F_{IS} global (moyenne sur l'ensemble des sous-échantillons et des loci) (estimation par f de Weir et Cockerham, 1984). Ce processus est répété un très grand nombre de fois, ce qui permet d'obtenir la distribution des F_{IS} générés sous l'hypothèse de panmixie locale (H_0). Trois tests sont ensuite possibles (en toute rigueur, il faut choisir lequel avant).

Tester s'il existe un déficit en hétérozygotes

Il s'agit donc d'un test unilatéral avec H_1 : le F_{IS} de la population échantillonnée est plus grand que 0. On teste bien ici si les sous-populations échantillonnées sont panmictiques (H_0) et n'utilisent pas un mode fermé de reproduction (du type autofécondation ou croisements entre apparentés) qui doit donner une homozygotie supérieure à celle attendue sous panmixie à tous les loci. La proportion de fois que l'on obtient, au cours de la randomisation, une valeur aussi grande ou plus grande que celle observée nous donne la *P*-value du test. Si le test est significatif, on peut ensuite essayer d'estimer le taux d'autofécondation ou de croisements frère-sœur qui permet d'expliquer le F_{IS} observé, ou encore s'il peut être expliqué par un effet Wahlund (voir plus loin).

Tester s'il existe un excès d'hétérozygotes

C'est le test unilatéral dans l'autre sens avec H_1 : le F_{IS} de la population est inférieur à 0. La proportion de fois que l'on obtient, dans les randomisations, une valeur aussi faible ou plus petite encore que celle observée nous donne la *P*-value du test. Ici, ce qui est recherché c'est si les sous-populations se reproduisent de façon asexuée (clonalité) stricte, auquel cas on attend un $F_{IS} < 0$ pour tous les loci.

Tester un écart dans n'importe quelle direction (excès ou déficit)

Il se peut également que l'on s'attende à ce que les différents sous-échantillons ou les différents loci répondent dans toutes les directions (configurations de tests bilatéraux). Dans ce cas, il y a deux problèmes. Le premier, assez simple à résoudre, consiste en l'obtention de la *P*-value bilatérale. Le second correspond à une décision statistique globale, car il y a en effet autant de *P*-values obtenues que de loci et/ou de sous-échantillons testés.

La *P*-value bilatérale s'obtient comme suit. Il faut faire les deux tests unilatéraux, ce qui fournit deux *P*-values. Soit $P_{\min}$ la plus petite de ces deux probabilités (test

unilatéral le plus puissant) et P_{max} la plus grande des deux (test unilatéral le moins puissant des deux). Alors, la P-value bilatérale est simplement égale à P_{min} + (1 - P_{max}). Comme mentionné plus haut, il se peut que P_{max} soit inutilisable ou bien qu'elle ne soit pas calculable. Dans ce cas, on peut aussi multiplier P_{min} par deux.

Pour le second problème, celui-ci est relatif aux problèmes liés aux tests multiples. Nous traiterons cela plus loin.

Autres méthodes pour tester l'écart à la panmixie

Tests exacts

Tester la conformité à la panmixie est synonyme de tester une conformité à une rencontre aléatoire des gamètes et donc aux proportions génotypiques attendues sous les hypothèses de Hardy-Weinberg (de la forme p_i^2, $2p_i(1 - p_i)$ pour un allèle i quelconque). Ceci peut être également testé avec le test exact de Haldane (1954) (souvent improprement appelé test exact de Fisher qui correspond en fait à autre chose) ou par la généralisation de ce test pour plus de deux allèles (Guo et Thompson, 1992), comme cela est proposé dans Genepop (Raymond et Rousset, 1995b). Je n'encourage cependant pas l'utilisation de ce test pour plusieurs raisons. D'abord, ce test analyse la distribution de tous les génotypes possibles et pas forcément ce qui est affecté par un système de reproduction particulier (telle que l'autofécondation). Pour un locus à plus de deux allèles, l'hypothèse nulle peut être rejetée parce que certaines classes génotypiques manquent au profit de certaines autres, alors que le reste est conforme à l'attendu sous panmixie. Un tel résultat sera difficile à interpréter biologiquement. Ensuite, ce test n'est réalisable que dans chaque sous-échantillon et pour chaque locus séparément. Il n'y a pas de test global possible et on se retrouve obligé de gérer une multitude de P-values, exactes certes, mais, en addition du premier problème, bien difficile à exploiter en termes d'inférence biologique.

Méthode de Rousset et Raymond (1995)

Dans le logiciel Genepop, ce n'est pas f de Weir et Cockerham (1984) mais c'est un analogue de l'estimateur de Robertson et Hill (1984) qui est utilisé comme statistique pour effectuer le test. Comme nous l'avons déjà évoqué, cet estimateur est biaisé mais montre des variances en général plus faibles (pour des valeurs faibles). Les deux techniques convergent dans la plupart des situations et les différences obtenues ne concernent en général que les résultats à un locus dans un ou quelques sous-échantillons et n'ont donc que très rarement une conséquence interprétative (ou inférentielle) importante. Par conséquent, les deux procédures donnent presque toujours des résultats comparables.

Tester la pangamie

Tous les tests décrits précédemment ne sont en fait que des approches indirectes, puisque ce n'est que la conséquence de la rencontre aléatoire des gamètes qui est

testée, mais pas cette rencontre à proprement parler. Dans certaines circonstances, il est possible de tester la panmixie plus directement, si on a accès aux couples naturellement formés dans le milieu. En génotypant les adultes accouplés (en copulation), on peut tester si ces adultes se sont associés indépendamment de leurs génotypes, c'est-à-dire on peut tester la pangamie. Pour ce faire, il suffit d'adapter un test de MANTEL (1967), test que nous détaillerons davantage plus loin pour les tests de corrélations entre matrices de distances, pour tester la corrélation entre la matrice des distances génétiques (apparentement) entre les individus possibles (entre les différents individus accouplés) et la matrice d'accouplement (en codant 0 pour les paires d'individus non accouplés et 1 pour les paires d'individus effectivement trouvés accouplés). Pour des organismes à sexes séparés, les matrices concernent les femelles d'un côté contre les mâles de l'autre. Attention, ce test de Mantel ne peut être effectué par Genepop qui ne gère que des demi-matrices en excluant les valeurs diagonales (dont on a besoin ici). Il faut donc effectuer le test avec un logiciel qui utilise des données en colonne (comme Fstat, ou RT de MANLY, 1997). En permutant les cases d'une des deux matrices et en calculant un coefficient de corrélation à chaque fois, on obtient ensuite la probabilité d'observer une valeur aussi extrême ou plus extrême que celle observée. Cette procédure, malgré son intérêt évident, n'a à notre connaissance été utilisée qu'à deux reprises : chez le trématode *Schistosoma mansoni* chez les rats de Guadeloupe (PRUGNOLLE *et al.*, 2004b) et chez la tique du bétail *Rhipicephalus* (*Boophilus*) *microplus* en Nouvelle-Calédonie (CHEVILLON *et al.*, 2007a). Dans le premier cas, l'apparentement entre les paires d'individus a été effectué à l'aide du logiciel Kinship V.1.2. (module Relatedness) développé par K. F. Goodnight (http://gsoft.smu.edu/GSoft.html) qui calcule un estimateur d'apparentement non biaisé équivalent de celui décrit dans QUELLER et GOODNIGHT (1989). Pour les tiques, c'est l'estimateur de WANG (2002) qui a été préféré, car particulièrement robuste aux petits échantillons. Ce dernier fut calculé par le logiciel MER V3 (http://www.zoo.cam.ac.uk/ioz/software.htm#MER). Nous reverrons ce dernier exemple dans la seconde partie de ce manuel.

Dans tous les cas, l'information apportée par ce test peut s'avérer précieuse pour discuter des hypothèses possibles en vue d'expliquer une déviation du F_{IS} par rapport aux attendus sous panmixie.

TESTER LA STRUCTURATION

Tester le F_{ST}

Il s'agit de simuler la migration libre des individus entre sous-échantillons (H_0) en redistribuant au hasard les individus dans ces différents sous-échantillons. On mesure alors le F_{ST} obtenu avec θ (sur l'ensemble des loci). La répétition de ce

processus un très grand nombre de fois (10 000) nous permet d'obtenir une distribution des F_{ST} possibles sous H_0. L'hypothèse alternative H_1 correspond nécessairement à : « Il y a structuration », ou autrement dit, « le F_{ST} de l'échantillon est plus grand que 0 ». La *P*-value est donc donnée par la proportion de cas où le F_{ST} simulé a été aussi grand ou plus grand que le F_{ST} observé. Ce test est disponible dans Genetix qui propose également la même procédure avec le θ_{RH} de Robertson et Hill (1984) et le θ_{RH}' de Raufaste et Bonhomme (2000).

La méthode basée sur le *G* de Goudet *et al.* (1996)

En fait, certains travaux ont montré que le calcul d'une autre statistique (*G*) permettait d'avoir une plus grande puissance du test dans la plupart des situations (voir Goudet *et al.*, 1996). La procédure est rigoureusement identique sauf que l'on mesure un *G* (logarithme népérien de la vraisemblance du tableau de contingence observé) au lieu d'un F_{ST}. Cette statistique est calculée à partir d'effectifs alléliques, mais ce sont bien les individus diploïdes qui sont permutés au cours des randomisations (d'où le qualificatif de test génotypique). Une description de la formule du *G* peut être trouvée dans n'importe quel ouvrage de statistiques (Sokal et Rohlf, 1981) (voir aussi la réponse 7). Un avantage supplémentaire de cette statistique concerne ses propriétés additives, ce qui autorise la mise en place d'un test global sur l'ensemble des loci, comme on le retrouve dans Fstat.

$$G = -2\sum_{l=1}^{nl}\sum_{k=1}^{np}\sum_{i=1}^{na} N_{ikl} \ln\left(\frac{N_{ikl}}{N_{kl}\,\overline{p}_{il}}\right) \quad (46)$$

où *l* indique le locus et *nl* est le nombre total de loci, *k* les sous-échantillons et *np* le nombre total de sous-échantillons, *i* l'allèle et *na* le nombre total d'allèles au locus *l* dans la population *k*, N_{ikl} est le nombre de fois que l'allèle *i* du locus *k* est rencontré dans la population *l*, N_{kl} est le nombre d'allèles (deux fois la taille du sous-échantillon chez des diploïdes) du locus *l* dans le sous-échantillon *k* et $\overline{p}_{il}$ est la fréquence moyenne de l'allèle *i* du locus *l* dans tout l'échantillon. C'est donc cette statistique qui est calculée sur les données observées et pour chaque randomisation des individus entre sous-échantillons.

Test exact allélique de Rousset et Raymond (1995)

Il existe une autre solution pour tester la différenciation entre dèmes, mise au point par Raymond et Rousset (1995a). Il s'agit d'un test purement allélique qui fait donc l'hypothèse d'une indépendance totale des allèles dans les individus (panmixie parfaite). Pour que ce test soit valide, il est donc indispensable que les génotypes soient en parfaite conformité avec les attendus sous Hardy-Weinberg, car ce sont les allèles qui sont ici randomisés entre sous-échantillons. Ce test est proposé comme test allélique dans Genepop. Une procédure équivalente, le test « assuming HW » est

proposé dans Fstat (qui utilise un test basé sur le *G*). C'est le test le plus puissant qui existe, mais, parce qu'il est probable qu'aucune population ne soit en conformité avec une panmixie parfaite, je conseillerai de ne jamais appliquer ces procédures et de leur préférer celles utilisant les génotypes (ne supposant donc pas la panmixie). Par ailleurs, le test exact ne peut être effectué que locus par locus, ce qui impose une procédure supplémentaire pour obtenir un test global (voir plus loin le paragraphe sur les tests multiples). Ajoutons enfin que les logiciels cités traitent les données haploïdes en dédoublant chaque allèle (homozygotie artificielle totale). Dans ce cas, le test allélique est impossible ou alors doit être fait sur la moitié des individus si c'est possible.

TESTER LA PANMIXIE GLOBALE

Ceci est fait en réassociant au hasard les allèles des individus de l'ensemble de l'échantillon un très grand nombre de fois. On mesure le F_{IT} sur l'ensemble des loci. Pour le reste, la procédure est identique à celle présentée pour tester la significativité du F_{IS}.

Il peut sembler redondant de tester le F_{IT} après avoir testé le F_{IS} et le F_{ST}, mais dans certains cas cela peut s'avérer utile. En particulier, un F_{IT} nul associé à d'autres critères (voir plus loin) peut être diagnostique d'une espèce strictement clonale et fortement structurée en de nombreux dèmes (voir De Meeûs et Balloux, 2005 ; Nébavi *et al.*, 2006).

TESTER LES DÉSÉQUILIBRES DE LIAISON

Ici, plusieurs méthodes sont possibles. Globalement, elles consistent à recombiner au hasard les loci entre eux à l'intérieur de chaque sous-échantillon un très grand nombre de fois et de mesurer (différentes méthodes) une statistique. La statistique observée dans chaque sous-échantillon est ensuite comparée à la distribution obtenue lors des randomisations sous l'hypothèse nulle d'absence d'association statistique entre loci. La statistique peut être une mesure de déséquilibre de liaison par paire de loci (le plus fréquent), ou une mesure multiloci (utilisée par les chercheurs travaillant sur des organismes clonaux). L'avantage des mesures multiloci est qu'elles fournissent une mesure sur l'ensemble des loci, alors qu'il y a autant de mesures (et donc de tests) qu'il y a de paires de loci (potentiellement $L(L-1)/2$ où L est le nombre

de loci) pour les mesures par paire. Le défaut des mesures multiloci est que leur comportement n'est pas encore bien connu dans toutes les conditions (voir De Meeûs et Balloux, 2004) et qu'il n'existe pas de mesure (et donc de test) multi-échantillons. Dans les tests par paire de loci, on peut utiliser comme statistique la probabilité d'apparition du tableau des génotypes pour les deux loci du sous-échantillon, compte tenu des fréquences génotypiques observées. Dans ce cas, la *P*-value du test sera simplement la somme des probabilités aussi faibles ou plus faibles que celle observée dans le sous-échantillon (voir le système d'équations 25), divisée par la somme de toutes les probabilités obtenues lors de la procédure de randomisation. Autrement dit, si P_{obs} est la probabilité du tableau des génotypes observés pour la paire de loci L1_L2 dans le sous-échantillon S1, P_i la probabilité d'occurrence d'un tableau randomisé et *Rand* le nombre total de randomisations (nombre de fois que les génotypes ont été recombinés librement), alors la *P*-value du test de déséquilibre de liaison sera :

$$P = \frac{\sum_{i=1}^{i=Rand} \left(P_i \leq P_{obs} \right)}{\sum_{i=1}^{i=Rand} P_i} \qquad (47)$$

C'est ce qui est fait dans Genepop 3.4. (Raymond et Rousset, 2003, mis à jour de Raymond et Rousset, 1995b). On peut aussi calculer une autre statistique, telle qu'un *G* comme dans le logiciel Fstat 2.9.3. (Goudet, 2002, mise à jour de Goudet, 1995) et Genepop 4 (Rousset, 2008), ou sur un coefficient de corrélation comme dans Genetix 4.03 (Belkhir *et al.*, 2004) ou encore sur un estimateur multilocus comme dans Multilocus 1.3b (Agapow et Burt, 2003, mis à jour d'Agapow et Burt, 2001).

À partir d'ici, plusieurs points importants doivent être précisés.

Nombre de randomisations

Certaines procédures de randomisations peuvent être très gourmandes en nombre de randomisations. Ce nombre sera fonction du nombre de combinaisons de génotypes possibles entre les deux loci étudiés. Dans le doute, il faut donc bien veiller à vérifier que deux procédures de randomisations faites indépendamment sur les mêmes données donnent le même résultat. Ceci est particulièrement important pour la procédure (chaîne de Markhov) utilisée dans Genepop où le nombre d'itérations devra atteindre au moins 10^6, voire 10^7.

Correction du seuil

Comme nous l'avons vu, les tests par paire de loci génèrent un grand nombre de tests (autant que de paires de loci). Pour sept loci, par exemple, on a 21 paires de loci

possibles. Cette répétition de tests va poser un problème statistique important que nous traiterons dans la section suivante. Ces tests sont par ailleurs non indépendants puisque chaque locus est comparé à chacun des autres loci restants, ce qui signifie que l'information contenue dans chaque locus est utilisée de façon redondante, ce qui pose un problème supplémentaire. Dans le paragraphe qui suit, nous verrons comment corriger le seuil de décision statistique afin de prendre en compte ces difficultés.

Remarques sur les tests de déséquilibres de liaison et leur interprétation

Comme nous l'avons déjà vu, certaines des procédures que nous utilisons en génétique des populations empiriques requièrent l'utilisation d'un nombre important de loci (au moins cinq) qui devraient être indépendants statistiquement. C'est-à-dire que l'information portée par chacun de ces loci est supposée indépendante. Un déséquilibre de liaison fort risquerait d'apporter une redondance forte conduisant à un risque d'erreur de décision. En fait, l'indépendance des loci ne peut être certaine que si les populations échantillonnées sont de tailles infinies, panmictiques et non structurées et ce depuis un grand nombre de générations, ce qui n'est évidemment jamais le cas. Il y a donc toujours liaison. Le principal est que cette liaison ne nuise pas trop à la détection du signal recherché. Le reste est laissé à l'appréciation de chacun, mais fort heureusement ces tests sont individuellement peu puissants et les procédures qui y sont le plus souvent associées (Bonferroni) rendent la détection de tels déséquilibres peu fréquente. De ma propre expérience sur les populations clonales (déséquilibres de liaison forts à totaux), c'est plus une diminution de puissance des tests (de différenciation, en particulier) qu'une augmentation qu'il faut attendre (augmentation des variances d'estimation), comme cela peut être illustré par les immenses intervalles de confiance de F_{ST} obtenus par bootstrap sur les loci chez la levure opportuniste *Candida albicans* (voir la figure 1 dans Nébavi *et al.*, 2006).

LE PROBLÈME DES TESTS RÉPÉTÉS

Comme nous l'avons déjà vu, le but d'un test statistique est d'évaluer la probabilité avec laquelle le hasard permet d'expliquer nos données si celles-ci proviennent d'une population respectant l'hypothèse nulle. Si cette probabilité est inférieure à un seuil choisi α, on décide que les données dévient significativement de ce que l'on attend sous H_0. Par conséquent, et par définition, pour un seuil choisi de $\alpha = 0{,}05$ (le plus classique), on s'attend à ce que sous H_0 5 % des tests soient significatifs par hasard. Autrement dit, si j'échantillonne 100 fois dans une population panmictique et que

j'effectue un test du F_{IS} pour chaque échantillon, je m'attends à trouver en moyenne cinq tests significatifs au seuil $\alpha = 0,05$ (si la taille des échantillons et si le polymorphisme des loci sont suffisants).

Par conséquent, la répétition de tests pose un problème. Plusieurs méthodes existent pour résoudre le problème des tests répétés et dont l'application dépend de la question posée et du type de tests répétés.

Les tests répétés sont indépendants

Ces tests répétés peuvent correspondre à différents cas de figure dont voici une liste non exhaustive :

– je voudrais combiner différents tests (de la même H_0) trouvés dans la littérature pour lesquels je n'ai pas les données brutes ;

– je cherche à savoir si le F_{IS} de chaque locus dévie significativement de 0 dans un sens ou dans l'autre ;

– je dispose de données de structuration de plusieurs sites comparables, sur plusieurs années et je cherche à combiner les *P*-values obtenues lorsque j'ai testé la significativité du F_{ST} dans chacun de ces jeux de données d'années différentes ;

– je compare la différenciation entre deux catégories d'individus (mâles *versus* femelles ; parasites d'hôtes d'espèces différentes ou de sexes différents, etc.) dans plusieurs sites (je souhaite combiner l'information de tous les sites).

Dans tous les cas, je peux chercher à savoir si un signal global existe ou je peux désirer identifier quels tests sont significatifs.

Tester si un signal global existe

On peut alors combiner les *k* tests de quatre façons différentes : le test binomial et sa version généralisée (Teriokhin *et al.*, 2007 ; De Meeûs *et al.*, 2009), la procédure de Fisher (Fisher, 1970), le test SGM (Goudet, 1999) et la transformation *Z* de Stouffer (Whitlock, 2005).

On peut procéder à un test binomial pour un nombre d'essais correspondant au nombre de tests et un attendu correspondant au seuil α. Pour $\alpha = 0,05$, la structure du test est la suivante :

– H_0 : la proportion de tests significatifs observés n'est pas différente de 0,05 ;

– H_1 : la proportion observée de tests significatifs est supérieure à l'attendu 0,05 (test unilatéral).

La plupart des logiciels de statistiques font le test binomial et son application est assez simple. La loi binomiale concerne les cas où on ne peut avoir que deux possibilités : vrai ou faux, présence ou absence, noir ou blanc ou, comme ici, significatif ou non. Elle est définie par le nombre d'essais (ou taille de l'échantillon) *k*, les probabilités

complémentaires d'état de l'événement *p* et *q* = 1 - *p* pour significatif et non significatif respectivement et *k'* le nombre de fois où l'événement « significatif » a effectivement été observé parmi les *N* essais. Dans notre cas, *k* correspond donc au nombre de tests que l'on souhaite combiner, et *k'* au nombre de tests significatifs au seuil de 5 % parmi ces *k* tests. On souhaite avoir la probabilité d'obtenir par hasard un nombre de tests significatifs aussi grand ou plus grand que *k'*. Cette probabilité est :

$$P = \sum_{i=k'}^{i=k} \frac{k!}{i!\,(k-i)!}\alpha^{i}(1-\alpha)^{(k-i)} \tag{48}$$

où $k! = k(k - 1)(k - 2)\ldots(k - k + 2)$

Donc si on a dix tests dont cinq sont significatifs, on a P = 0,00006 (valeur hautement significative donc). Pour un seul test significatif observé sur 10, cette *P*-value devient 0,4. Il existe depuis peu une version généralisée (Binomial généralisé) de ce test (Teriokhin *et al.*, 2007) implémentée par le logiciel MultiLocus V2.2 (De Meeûs *et al.*, 2009). La philosophie de ce test est décrite en détail dans l'aide qui accompagne le logiciel et je ne reviendrai donc pas dessus.

La procédure de Fisher (Fisher, 1970), qu'il ne faut pas confondre avec le test exact du même auteur car cela n'a pas de rapport, propose la formule suivante :

$$\chi^2_{obs} = -2\sum_{i=1}^{i=k} Log\ (P_i) \tag{49}$$

où P_i correspond à la *P*-value obtenue au *i*ème test.

Cette expression suit normalement une loi du χ^2 (Chi-2) avec 2*k* degrés de liberté (ddl), dont on peut donc extraire la *P*-value associée à partir d'une table du χ^2, d'un logiciel ou en tapant la formule LOI.CHIDEUX(χ^2_{obs};2*k) sous Excel.

Le test de randomisation SGM de symétrie autour de 0,5 de la moyenne géométrique (la moyenne géométrique correspond à la racine *k*ième du produit des *k* *P*-values entre elles) (Goudet, 1999) est implémenté par le logiciel SGM distribué sur demande par l'auteur lui-même.

Le test de transformation *Z* de Stouffer (Whitlock, 2005) consiste en la transformation des *P*-values en leur équivalent *Z*, avec par exemple la commande Excel LOI.NORMALE.INVERSE(P_i;0;1) ou LOI.NORMALE.STANDARD.INVERSE(P_i) (mettre 0,9999 pour les P_i = 1) qui donne un Z_i pour chaque P_i que l'on combine en la statistique Z_s :

$$Z_s = \frac{\sum_i^k Z_i}{\sqrt{k}} \tag{50}$$

La *P*-value globale est obtenue en comparant cette statistique à la loi normale, avec par exemple la commande Excel LOI.NORMALE.STANDARD(Z_s).

D'une façon générale, si on combine peu de tests ($k < 4$) il vaut mieux appliquer le test *Z* de Stouffer et, dans les autres cas ($k > 3$), le binomial généralisé ou le *Z*

(De Meeûs *et al.*, 2009). Par ailleurs, d'une façon qui ne concerne pas vraiment notre propos ici, le test binomial n'exige pas de connaître la *P*-value exacte des tests à combiner (même si cela est préférable), ce qui peut représenter un avantage certain lorsque l'on combine des données de la littérature.

On pourra trouver une discussion plus théorique de ce type de problèmes dans la littérature (Goudet, 1999 ; Whitlock, 2005 ; De Meeûs *et al.*, 2009). La procédure de Fisher sera adéquate dans certaines configurations de distribution des *P*-values (en U, en cloche, en L, ou en J) autres que la distribution uniforme. Il est en effet important de faire attention à cela et ne pas être esclave de ses données. La présence d'une *P*-value = 0 devrait en toute rigueur interdire l'utilisation de la procédure de Fisher.

La procédure de Fisher répond davantage à la question : y a-t-il au moins un test significatif ?

Le SGM est quant à lui très (trop) conservateur, une propriété qui pourrait s'avérer utile dans le cadre des méta-analyses (sur jeux de données publiées) où le biais de publication en faveur des résultats significatifs pourrait être ainsi partiellement corrigé.

Déterminer quels sont les tests significatifs, procédure de Bonferroni

Une autre configuration pourrait nous amener à rechercher lesquels, parmi ces *k* tests, sont réellement significatifs. Ce peut être le cas si on recherche un marqueur de sous-dominance au milieu de plusieurs marqueurs (quels loci sont déficitaires en hétérozygotes ?). Dans ce cas, il n'y a pas d'autre solution que de procéder à une correction de Bonferroni (très conservatrice) (Holm, 1979 ; Rice, 1989) ou, de façon moins conservatrice (sauf pour la plus basse *P*-value), le Bonferroni séquentiel.

Il faut ordonner les *N* *P*-values de chaque test de la plus petite à la plus grande. La plus petite des probabilités est multipliée par *N*, la deuxième plus petite par $N - 1$, la troisième par $N - 2$, etc. Les tests significatifs sont ceux dont la *P*-value ainsi corrigée reste inférieure au seuil choisi α (= 0,05). On peut aussi plus classiquement corriger le seuil lui-même par le nombre de tests correspondants : $\alpha'_1 = \alpha/N$, $\alpha'_2 = \alpha/(N - 1)$, $\alpha'_3 = \alpha/(N - 2)$, etc., ce qui revient au même, mais personnellement je préfère disposer d'une *P*-value « exacte ». Le tableau 4 donne un exemple d'une série de 10 tests triés par ordre croissant de leur *P*-value, leur *P*-value corrigée par la procédure du Bonferroni séquentiel, le seuil corrigé correspondant à $\alpha = 0,05$, ainsi que la décision statistique concernant chaque test, compte tenu du nombre de tests effectués.

Dans cet exemple on décide que trois des dix tests ont donné une valeur déviant significativement de ce qui est attendu sous H_0. On remarquera que cette procédure est très conservatrice. Il faut le savoir au moment d'échantillonner et ne pas lésiner

Tableau 4
Exemple d'application de la procédure du Bonferroni séquentiel sur un jeu de 10 tests. Les *P*-values ont été classées par ordre croissant.

Test N°	*P*-value	Nombre de tests	*P*-value corrigée	α'	Décision
2	0,001	10	0,010	0,0050	**
9	0,003	9	0,027	0,0056	*
10	0,005	8	0,040	0,0063	*
8	0,015	7	0,105	0,0071	ns
7	0,022	6	0,132	0,0083	ns
4	0,041	5	0,205	0,0100	ns
3	0,050	4	0,200	0,0125	ns
1	0,101	3	0,303	0,0167	ns
6	0,210	2	0,420	0,0250	ns
5	0,321	1	0,321	0,0500	ns

** : significatif au seuil 1 %, * : significatif au seuil 5 %, ns : non significatif.

sur le nombre d'individus génotypés. De faibles échantillons ne permettront jamais d'obtenir des *P*-values suffisamment basses pour supporter une procédure de Bonferroni. C'est ce que les statisticiens appellent le risque de seconde espèce (β) ou risque de se tromper en acceptant H_0. Alors que α correspond au risque de première espèce, ou risque de se tromper en rejetant H_0. Nous avons déjà évoqué ces concepts en p. 69.

Les tests répétés ne sont pas indépendants

C'est typiquement le cas des tests de déséquilibre de liaison par paire de loci. C'est aussi le cas de tests de différenciation par paire de sous-échantillons. Ici encore, il est nécessaire de distinguer le cas où une réponse globale est souhaitée du cas où on recherche quels tests sont significatifs.

Tester si un signal global existe

Nous prendrons l'exemple des tests de déséquilibre de liaison par paire de loci dans la mesure où des tests globaux existent normalement dans les autres cas (différenciation). On peut alors appliquer le test binomial ou le Z. En cas de signal positif (H_1 vraie), il y aura autocorrélation entre les tests (si A et B sont liés ainsi que B et C, alors A et C seront liés) et donc inflation de la puissance des tests. La précision du

niveau de significativité n'aura donc pas beaucoup de sens, ce sera significatif ou pas. Comme sous l'hypothèse nulle, nous n'attendons pas plus de 5 % de tests significatifs à ce seuil et en principe pas d'autocorrélation entre tests, le test exact binomial doit faire l'affaire. Sous R (R-Development-core-team, 2011), la commande est binom.test (*k*', *k*, 0,05, alternative ="greater") où *k*' est le nombre de tests significatifs au seuil 5 % parmi les *k* tests.

Déterminer quels sont les tests significatifs, procédure de Bonferroni

Ici, s'il est souhaité de déterminer quelles paires de loci sont statistiquement associées ou quelles paires de sous-échantillons sont différenciées, la procédure du Bonferroni séquentiel s'impose.

Le cas des déséquilibres de liaison

Les déséquilibres de liaison représentent le cas de figure le plus fréquent. Ici, en fonction de la taille de chaque sous-échantillon et du degré de polymorphisme des loci, le seuil de Bonferroni peut s'avérer impossible à atteindre (trop conservateur). Dans ce cas, il est plus raisonnable de ne prendre en compte que les loci les plus polymorphes, c'est-à-dire ceux pour lesquels les chances de détecter quelque chose sont les plus grandes. Par exemple, on peut écarter les loci dont un allèle atteint ou dépasse la fréquence de 90 %, on peut même être plus sévère en fonction des circonstances, car de tels loci ne présenteront qu'exceptionnellement des *P*-values suffisamment faibles alors qu'ils contribuent à l'augmentation de la sévérité du Bonferroni. Comme suggéré précédemment, ce qui est la plupart du temps recherché c'est si les loci ne sont pas trop liés. Il n'y a de toutes les façons pas d'agrément général sur la meilleure procédure et il est donc laissé libre choix aux empiristes de décider si une correction plus ou moins sévère ou pas de correction doit être appliquée. Les tests Multilocus (Agapow et Burt, 2001) ont été spécifiquement conçus pour tester un effet global, tel que celui attendu sous régime clonal de reproduction. Des études de simulations (De Meeûs et Balloux, 2004) ont suggéré que la mesure la plus précise (ou plutôt la moins mauvaise) serait le coefficient de corrélation r_D (Agapow et Burt, 2001). Ce coefficient se base sur l'indice d'association I_A (Brown *et al.*, 1980 ; Maynard-Smith *et al.*, 1993 ; Haubold *et al.*, 1998), mais contrairement à ce dernier est indépendant du nombre de loci étudiés dans l'analyse. Cette mesure est également utilisée comme statistique dans les tests de randomisation implémentés dans Multilocus. Par rapport au test bi-locus cette procédure permet l'obtention directe d'un test global sur l'ensemble des loci, mais ne peut être réalisée que sous-échantillon par sous-échantillon. Il peut cependant n'être significatif qu'à cause d'une seule paire de loci. Le test bi-loci de Fstat (basé sur le *G*) permet d'obtenir un test sur l'ensemble des sous-échantillons, mais pour chaque paire de loci prise une à une.

TESTER LA CORRÉLATION ENTRE DISTANCES

Il s'agit ici de regarder si la différenciation génétique (distance génétique) que l'on observe entre les sous-populations de notre échantillon peut être due aux distances géographiques qui les séparent, ou à d'autres distances. Ces autres distances peuvent être écologiques (différences de températures moyennes de janvier, pluviométriques, etc.), ou même génétiques si on souhaite comparer les distances génétiques hôtes et parasites, par exemple.

Dans tous les cas, on cherche à corréler deux matrices de distances entre elles. Comme pour les déséquilibres de liaison, les mesures de ces matrices ne sont pas indépendantes, ce qui empêche de procéder à des tests classiques de corrélation ou de régression. Nous allons donc effectuer un test de MANTEL (1967).

La valeur d'un paramètre d'association, ou d'un coefficient de corrélation, entre les deux matrices est calculée à partir des données réelles, puis comparée à la série de pseudo-valeurs obtenues par permutation aléatoire de l'ordre des populations dans l'une des deux matrices de distances. À partir de là, la suite devient similaire à tout autre test par permutation. On pourra consulter la réponse 8 pour plus de détails sur le test de Mantel.

Distances génétiques et géographiques

Ce cas de figure a été étudié en profondeur par ROUSSET (1997). Ici, les sous-échantillons peuvent être distribués de deux façons différentes qui requièrent chacune une analyse qui lui est particulière. Cependant, dans les deux cas, la matrice des distances génétiques doit contenir une mesure corrigée de la différenciation entre paire de sous-populations, à savoir $\theta/(1 - \theta)$ (voir ROUSSET, 1997), θ étant l'estimateur du F_{ST} (voir p. 53). On sait en effet par l'équation (21) que :

$$F_{ST} = \frac{Q_S - Q_T}{1 - Q_T}$$

Sachant que, dans le cas d'un isolement par la distance, c'est-à-dire quand la différenciation augmente avec l'éloignement géographique des individus, et si chaque dème connaît un fonctionnement raisonnablement similaire (à peu près même taille et même système de reproduction) on voit bien que Q_S, la probabilité d'identité de gènes entre deux individus de la même sous-population, sera à peu près la même d'une sous-population à l'autre alors que Q_T, probabilité d'identité entre dèmes, sera une fonction décroissante de la distance entre dèmes. On voit donc bien que puisque Q_T se trouve au numérateur, mais aussi au dénominateur du F_{ST}, la relation entre F_{ST} et la distance géographique ne peut pas être linéaire. Par contre, comme on peut le voir, le rapport $F_{ST}/(1 - F_{ST})$ ne subit pas ce problème, en effet :

$$\frac{F_{ST}}{1-F_{ST}} = \frac{\dfrac{Q_S - Q_T}{1-Q_T}}{1-\dfrac{Q_S - Q_T}{1-Q_T}} = \frac{\dfrac{Q_S - Q_T}{1-Q_T}}{\dfrac{1-Q_T-Q_S+Q_T}{1-Q_T}} = \frac{Q_S - Q_T}{1-Q_S} \tag{51}$$

À partir de là, deux cas de figures sont à distinguer en fonction du schéma de dispersion des organismes étudiés, indépendamment du plan d'échantillonnage, qui peut suivre un tracé en une ou deux dimensions.

Les sous-échantillons sont alignés en une seule dimension

Certains organismes sont susceptibles de se distribuer en une seule dimension et seront donc échantillonnés comme tels. C'est typiquement ce qui se passe pour des organismes côtiers comme des bivalves (moules), des cirripèdes (balanes et anatifs), des patelles, des algues (fucus vésiculeux, laminaires), ou les organismes vivant le long des cours d'eau (mouches tsé-tsé, lymnées). La matrice des distances par paire de sous-échantillons est alors comparée à celle des $\theta/(1 - \theta)$ (car c'est bien l'estimateur non biaisé de Weir et Cockerham qu'il faut utiliser). Si le test de Mantel est significatif, on peut alors utiliser la pente b de la droite de régression du $F_{ST}/(1 - F_{ST}) = a + bD_G$, où D_G est la distance géographique, afin d'estimer le produit $D\sigma^2$ de la densité d'adultes reproducteurs dans un site (D) et la distance moyenne au carré mesurée entre la position occupée par un adulte reproducteur et celle que sa mère occupait à la génération précédente. En effet, Rousset (1997) montre qu'alors :

$$D\sigma^2 = \frac{1}{4b} \tag{52}$$

Cette méthodologie fut appliquée pour l'escargot intertidal (côtier) *Bendicium vitatum* (voir Rousset, 1997 pour une réanalyse) ou la tique d'oiseaux marins *Ixodes uriae* (McCoy *et al.*, 2003).

Les sous-échantillons sont distribués sur deux dimensions

Dans ce cas, le test de Mantel doit être effectué entre la matrice des Log népériens des distances géographiques par paire de populations et celle des $\theta/(1 - \theta)$ (voir Rousset, 1997). Si le test est significatif, la pente de la régression $F_{ST}/(1 - F_{ST}) \approx a + b\mathrm{Ln}(D_G)$ va permettre d'estimer $D\sigma^2$ avec l'équation (Rousset, 1997) :

$$D\sigma^2 = \frac{1}{4\pi b} \tag{53}$$

Si l'un des deux paramètres D ou σ peut être estimé, même approximativement, de façon indépendante, on obtient un pouvoir d'inférence relativement puissant ici (voir Koffi *et al.*, 2006a ; Bouyer *et al.*, 2009 ; De Garine-Wichatitsky *et al.*, 2009 pour illustration).

Les mêmes procédures peuvent être appliquées entre individus entre lesquels un équivalent du $F_{ST}/(1 - F_{ST})$ appelé a_r (calculé dans Genepop) et développé par Rousset (2000), Leblois *et al.* (2003) et Leblois *et al.* (2004) peut être régressé contre les distances entre individus (directe pour une dimension, en Log pour deux dimensions), ce qui conduit aux mêmes possibilités d'inférences que celles décrites ci-dessus. Watts *et al.* (2007) proposent une statistique *e* en principe plus puissante lorsque le voisinage ($4D\sigma^2$ ou $4\pi D\sigma^2$) est grand. Nous verrons cela plus en détail dans la partie pratique de ce manuel.

Dans le cas particulier de deux dimensions, Rousset (1997) montre que le nombre d'immigrants présents dans un sous-échantillon peut directement être tiré de la pente de la régression $F_{ST}/(1 - F_{ST}) \approx a + b\text{Ln}(D_G)$, $Nm = 1/2\pi b$.

Autres distances

On peut souhaiter vérifier si la différenciation entre sites est corrélée à une différence écologique entre sites ou tester s'il existe une corrélation entre différenciation génétique des sous-échantillons des hôtes et des parasites qui les infestent. Comme nous l'avons vu précédemment, le F_{ST} a été défini dans le cadre d'un modèle en îles. De fait, il ne se comporte pas idéalement par paire de populations (fortes variances, voir Balloux et Goudet, 2002) et on lui préférera d'autres mesures pour les tests de Mantel telles que la distance harmonique (*chord distance*) de Cavalli-Sforza et Edwards (Cavalli-Sforza et Edwards, 1967) ou la distance d'allèles partagés (*shared allelic distance*) (Bowcock *et al.*, 1994) (déjà discuté en p. 60). Pour la construction d'arbres (dendrogrammes), il semble aussi que les distances harmoniques donnent de meilleurs résultats (Takezaki et Nei, 1996).

En fait, la performance de différentes mesures et leur choix vont dépendre des situations rencontrées, même si en principe toutes les distances devraient aboutir en théorie à des résultats concordants. Ceci peut être illustré par la corrélation que Prugnolle *et al.* (2005) ont montrée entre les distances génétiques entre infra-populations[8] de schistosomes et celles mesurées entre les rats qui les portaient (ou leur apparentement si on préfère) en Guadeloupe. Dans l'article, c'étaient la distance de Cavalli-Sforza et Edwards (1967) qui avait été utilisée entre infra-populations de schistosomes et la « *shared allele distance* » (Bowcock *et al.*, 1994) entre les individus rats. Le logiciel MSA (Dieringer et Schlötterer, 2003, téléchargeable à http://i122server.vu-wien.ac.at/) calcule cette distance. La corrélation obtenue était très significative (*P*-value = 0,0005), mais De Meeûs *et al.* (2007a) ont montré que si le F_{ST} est utilisé pour les deux matrices, la corrélation n'est plus significative (*P*-value = 0,15) et elle l'est beaucoup moins (*P*-value = 0,0113) quand c'est Cavalli-Sforza et Edwards qui est utilisé pour les deux matrices. Le choix d'une statistique n'est donc pas entièrement neutre. Ajoutons enfin que d'autres mesures d'apparentement entre individus existent,

[8] En parasitologie, une infra-population est le contenu en parasites d'un individu hôte.

telles que l'estimateur de QUELLER et GOODNIGHT (1989) ou de WANG (2002) dont nous avons déjà parlé à propos des tests de pangamie (p. 77-78) ou, plus récemment, de KALINOWSKI *et al.* (2006) qui pourrait être encore plus puissant.

TESTER LES BIAIS DE DISPERSION DE CERTAINES CATÉGORIES D'INDIVIDUS

Dans les populations naturelles, il se peut qu'un sexe disperse davantage que l'autre sexe ou que les individus parasités dispersent plus ou moins bien que les individus sains. Dans ce cas, il existe plusieurs statistiques (mesures) qui peuvent être comparées (GOUDET *et al.*, 2002). Je ne parlerai ici que de trois d'entre elles et dans le cas d'un biais de dispersion sexe-spécifique.

L'indice d'assignement, dont nous avons déjà parlé en p. 64, consiste à calculer la probabilité qu'un individu a d'appartenir à la sous-population où il a été échantillonné, compte tenu de son génotype à tous les loci génotypés et de celui de l'ensemble des individus de son sous-échantillon. Afin de tenir compte du degré de polymorphisme qui peut beaucoup varier d'un site à l'autre, il faut corriger cette probabilité. Ceci est fait en soustrayant à cette valeur la valeur moyenne obtenue sur l'ensemble des individus du sous-échantillon concerné, après une transformation Log afin de minimiser les risques d'erreurs dus aux petites valeurs (voir FAVRE *et al*,. 1997 pour plus de détails). Cet indice se note AI_c. Sa distribution sur l'ensemble des populations est nécessairement centrée sur 0. Et une valeur négative signifie que l'individu est moins bien assigné que la moyenne à son propre sous-échantillon. On calcule ensuite la moyenne de cet indice sur les mâles et la moyenne sur les femelles de l'ensemble de l'échantillon. La statistique suivante est calculée :

$$t = \frac{\overline{AI_c^-} - \overline{AI_c^+}}{\sqrt{\frac{s^2(AI_c^-)}{Nb(+)} + \frac{s^2(AI_c^+)}{Nb(-)}}} \qquad (54)$$

où les signes – et + désignent la catégorie qui disperse le moins et le plus respectivement, la barre désignant la moyenne, s^2 la variance et Nb le nombre total d'individus de la catégorie considérée, observés dans l'ensemble des sous-échantillons.

La moyenne du sexe le moins dispersant (donc mieux assigné) doit être supérieure à celle du sexe le plus dispersant.

La deuxième statistique qui nous intéresse correspond à :

$$Rs^2(AI_c) = \frac{s^2(AI_c^+)}{s^2(AI_c^-)} \qquad (55)$$

La variance de l'indice d'assignement du sexe le plus dispersant doit être supérieure à celle du sexe le moins dispersant.

La troisième statistique dépend de la différence des F_{ST} estimés pour chaque catégorie :

$$\Delta\theta = \theta(-) - \theta(+) \qquad (56)$$

La différenciation mesurée sur la catégorie d'individus les moins dispersants doit être plus élevée que celle mesurée pour la catégorie la plus vagile.

Ensuite, l'appartenance à une catégorie (mâle ou femelle) est re-distribuée au hasard pour chaque individu de chaque sous-échantillon, en gardant les individus dans leur sous-échantillon, et en conservant la même proportion de chaque catégorie (même sexe-ratio) et la statistique est mesurée. Cette randomisation est répétée un grand nombre de fois afin d'obtenir une distribution des valeurs possibles sous H_0 (pas de différence de dispersion) à laquelle la valeur observée est comparée. Les tests peuvent être unilatéraux ou bilatéraux. Dans ce dernier cas, ce sont les valeurs absolues des différences [dans (53) et (55)] ou le ratio de la plus grande sur la plus petite valeur de chaque randomisation qui sont utilisés. Ces mesures et randomisations sont toutes implémentées dans Fstat (menu "biased dispersal"). Ces procédures ont été utilisées avec succès pour mettre en évidence, dans les populations suisses de la tique *Ixodes ricinus*, un biais de dispersion sexe-spécifique, les femelles représentant le sexe peu ou pas dispersant (De Meeûs *et al.*, 2002a), et un biais de dispersion pathogène spécifique, les tiques infectées par le spirochète *Borrelia afzelii* dispersant très peu ou pas du tout (De Meeûs *et al.*, 2004b). De même, Prugnolle *et al.* (2002) ont pu mettre en évidence une structure génétique spécifique du sexe chez le trématode *Schistosoma mansoni* infectant des rats en Guadeloupe.

Dans certains cas, l'échantillonnage ne permet pas de tester une différence entre sexes ou entre catégories d'individus, par randomisation, notamment pour tester une différence de F_{ST}. Dans ce cas, une alternative moins puissante existe et permet de comparer H_s, F_{IS} ou le déséquilibre de liaison entre catégories d'individus dans un seul échantillon. Il suffit d'utiliser les loci (ou les paires de loci pour les déséquilibres de liaison) comme des répliquats (plus ou moins indépendants d'ailleurs) et de faire un test de comparaison pour données appariées, le critère d'appariement correspondant donc au locus (ou la paire de loci). Comme la distribution de telles données a toutes les chances de ne pas suivre une loi normale, il est conseillé ici de procéder à un test de rang de Wilcoxon pour données appariées (*Wilcoxon signed ranks test for paired data*) (Siegel et Castellan, 1988).

TESTER LA DIFFÉRENCE ENTRE GROUPES

Ce cas de figure se présente lorsque différents types de sites doivent être comparés. C'est typiquement le cas si on souhaite comparer différents paramètres génétiques,

tels que H_s, F_{IS}, F_{ST} ou d'autres, entre infra-populations trouvées dans des hôtes mâles et celles trouvées dans des hôtes femelles. Ce peut être aussi le cas entre des sites de différentes natures tels que des prés et des bois dans un paysage de bocages (par exemple, H_0 : les populations de bois sont-elles plus structurées que celles de pré, ou plus pamictiques, etc.). Ce peut également être le cas pour comparer des parasites trouvés sur des espèces hôtes différentes. Les procédures suivent toujours la même philosophie. Ici, le paramètre d'intérêt est moyenné sur l'ensemble des sous-échantillons de chaque catégorie. Soit x_i cette valeur moyenne pour les sous-échantillons du groupe i. Pour un test unilatéral avec deux groupes, on calcule juste la différence (x_1-x_2) (x_1 étant la plus grande). Pour les autres cas de figure, la statistique utilisée sera :

$$\Delta S_x = \sum_{i=1}^{ng-1} \sum_{j=i+1}^{ng} \left(x_i - x_j\right)^2 \qquad (57)$$

où *ng* représente le nombre de groupes à comparer.

Ensuite, les échantillons de chaque groupe sont randomisés (permutations aléatoires des échantillons dans les différents groupes en gardant le nombre d'échantillon par groupe constant) un grand nombre de fois (10 000) et la statistique (x_1-x_2) ou celle définie en (57) est recalculée pour chaque randomisation. La valeur observée est ensuite comparée à la distribution des valeurs obtenues par randomisation, la *P*-value du test correspondant (encore une fois) à la proportion de fois qu'une valeur aussi grande ou plus grande a été observée au cours des randomisations. Cette procédure est implémentée dans Fstat (menu "Comparison among groups").

Comme précédemment, l'échantillonnage peut ne pas permettre de procéder à ce test sans qu'il soit pour autant impossible de tester des différences de F_{IS}, de H_s ou de déséquilibres de liaison. Ici aussi, les loci (ou paires de loci) peuvent être utilisés comme répliquats pour un test de rangs pour données appariées (voir par exemple Nébavi *et al.*, 2006).

ANALYSES MULTIVARIÉES

Les analyses multivariées permettent souvent une représentation didactique de l'organisation générale de la variabilité génétique globale des échantillons génotypés. Dans certains cas, ils permettent également des analyses statistiques et des inférences. Il en existe plusieurs types, de même nature, mais offrant des possibilités différentes.

Analyse factorielle des correspondances (AFC)

Cette analyse, introduite par Benzécri (1973), a été adaptée aux données génétiques diploïdes par She *et al.* (1987). L'AFC place chaque individu dans un hyper-espace

à *K* dimensions (*K* étant le nombre total d'allèles présents sur l'ensemble des loci) et les projette sur les plans définis par les axes orthogonaux (donc indépendants) expliquant le mieux la dispersion des points (même principe que celui d'une régression). Une mesure de la pertinence des axes ainsi définis est représentée par le pourcentage d'inertie de chaque axe. Comme il y a *K* axes, un axe représentant 100/*K* % d'inertie ne veut rien dire. L'inertie est donc proportionnelle non seulement à la quantité d'information que l'axe correspondant représente, mais est aussi fonction du nombre total d'axes (plus il y a d'axes et moins chaque axe peut avoir une très forte inertie). L'AFC est une procédure qui peut s'avérer utile pour classer les individus en fonction de leur proximité génétique.

Exemples

L'utilisation de l'AFC s'est avérée payante pour analyser la présence de trématodes parasites dans une zone d'hybridation de leur hôte (moule de bouchot, *Mytilus edulis*) avec une autre espèce (moule d'Espagne, *M. galloprovincialis*) incompatible pour le parasite (Coustau *et al.*, 1991) ou, de façon plus spectaculaire, dans le cas du monogène *Diplozoon gracile*, spécifique du poisson *Barbus meridionalis*, en zone d'hybridation avec *B. barbus*, un hôte moins favorable au parasite, comme présenté dans la figure 11.

Cette technique peut également être utilisée pour détecter une structure cachée dans un échantillon comme celle qui proviendrait d'un effet Wahlund (déficits en hétérozygotes à tous les loci non expliqués par le système de reproduction), comme cela a été réalisé dans Solano *et al.* (2000) (voir le paragraphe suivant). GENETIX 4.05.4 (développé par Belkhir *et al.* et téléchargeable gratuitement à http://www.univ-montp2.fr/~genetix/genetix/genetix.htm) offre une interface extrêmement conviviale, en français qui plus est (assez rare pour être souligné), pour produire des AFC en deux ou même trois dimensions (pas nécessairement les plus faciles à lire en ce qui me concerne).

Recommandations et astuces pour les utilisateurs de l'AFC

Quand on procède à une AFC (FCA ou FA en anglais), le programme génère différents fichiers tels que celui contenant les coordonnées des individus sur les différents axes. On peut être tenté d'utiliser ces coordonnées, qui sont donc des données ordinales continues issues de données qualitatives disjointes, pour procéder à des analyses de type analyse de variance (Anova) ou régression. Vérifier si les coordonnées des individus sur le premier axe de l'AFC sont expliquées plus ou moins bien par telle ou telle autre variable écologique peut en effet représenter une perspective séduisante. Je sais que beaucoup de personnes considèrent qu'il n'est pas valide de procéder à ce genre d'analyses à partir d'une AFC (alors qu'ils considèrent que cela est possible à partir d'une ACP, traitée plus loin) qui transforme des données

discrètes bornées (0, 1 ou 2) en données continues de distribution incertaine. Mon opinion à ce sujet est que si on peut s'en passer on évite les ennuis, mais sinon je ne vois pas vraiment où est le problème à partir du moment où certaines précautions sont prises, comme de vérifier la distribution des données avant de procéder à une Anova.

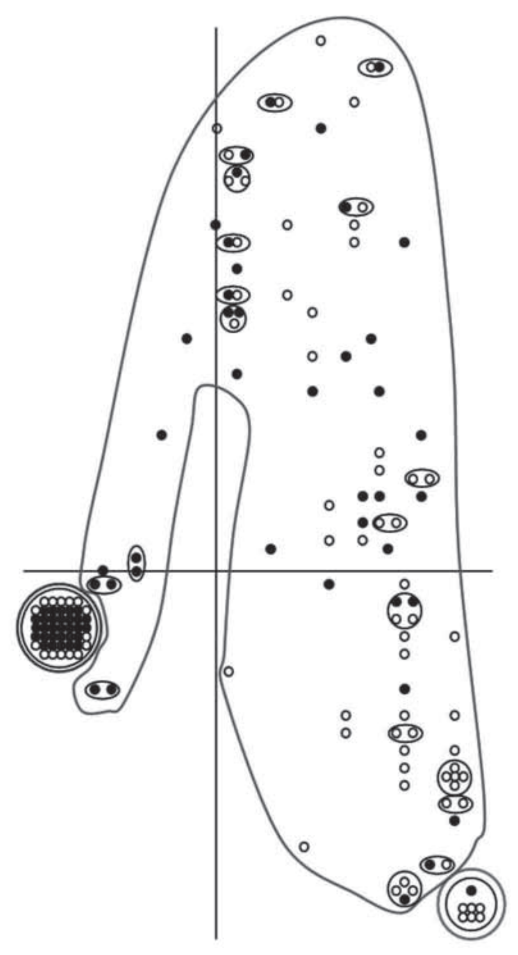


Figure 11
AFC d'individus hôtes *Barbus* sur le plan défini par les deux axes principaux de l'analyse, faite à partir de données sur neuf loci enzymatiques. Les génotypes *B. barbus* purs sont cerclés de rouge, les *B. meridionalis* purs sont cerclés de bleu et les hybrides de vert. Chaque rond correspond à un poisson, les ronds noirs étant les poissons parasités par *D. gracile*.
Les individus superposés (même coordonnées dans le plan) sont cerclés de noir. Le nuage de points en U inversé est typique de données changeant progressivement d'un état à un autre, comme les allèles dans une zone hybride, et s'appelle « Effet Guttman » (WOLFF, 1996).
On voit bien que les parasites suivent fidèlement cette forme en devenant de plus en plus fréquents au fur et à mesure que la fréquence des allèles de *B. meridionalis* augmente dans le génotype multilocus des individus hôtes (graphique tiré de DE MEEÛS *et al.*, 2007a).

Il existe aussi une astuce à connaître par rapport au fait que ce type d'analyse est très sensible à la présence d'individus porteurs d'un allèle rare (*outliers* en anglais). En effet, les individus porteurs d'un allèle rare vont tirer le nuage vers eux. Le résultat est néfaste, car les autres individus se retrouveront compactés dans un nuage trop dense pour qu'on puisse y détecter quoi que ce soit. Cela va aussi remettre sérieusement en cause toute utilisation des coordonnées, car les coordonnées de chaque individu seront alors conditionnées majoritairement par la position de quelques individus exceptionnels. Il est souvent nécessaire de retirer plusieurs individus de l'analyse et parfois même un grand nombre. Dans l'échantillon de Nyafaro (Burkina Faso) dans Solano *et al.* (2000), près de 42 % des individus ont dû être ainsi écartés de l'analyse afin de pouvoir déceler une sous-structure dans les individus restants (60 sur les 97).

Analyse en composantes principales (ACP)

Une ACP (PCA en anglais) suit le même principe que l'AFC sauf que ce sont des données ordinales continues qui sont utilisées au lieu de données disjonctives. Ici, ce sont des groupes d'individus (sous-échantillons) qui seront positionnés dans un hyperespace de K dimensions. Les coordonnées de chaque groupe sur chacun des axes principaux peuvent être utilisées pour des analyses statistiques supplémentaires telles que des analyses de variance ou autres régressions comme dans Nébavi *et al.* (2006). C'est une procédure fort utile pour positionner des sous-échantillons les uns par rapport aux autres en fonction de leur appartenance à un groupe écologique particulier comme des sous-échantillons de tiques d'oiseaux marins sur différentes espèces hôtes, comme on peut le voir dans la figure 12 (voir aussi McCoy *et al.*, 2003, 2005).

Le logiciel PCA-GEN ver. 1.2 (développé par J. Goudet librement téléchargeable à http://www2.unil.ch/popgen/softwares/pcagen.htm) permet cette analyse à partir de données au format Fstat (mais avec un format limité à deux caractères par allèle). Ce logiciel, en plus de fournir les graphiques en deux dimensions de la projection des points selon les axes demandés et leur pourcentage d'inertie, fournit également des tests de significativité de ces axes selon la méthode du bâton brisé (*broken stick*), une technique empirique appliquée à l'ACP (Frontier, 1976 ; Legendre et Legendre, 1998 ; King et Jackson, 1999) qui correspond davantage à un critère qu'à un test réel. Une explication plus détaillée de cette technique peut être consultée en réponse 9 à la fin de ce manuel. PCA-GEN propose aussi une procédure de permutations des génotypes complets entre sous-échantillons afin de tester la significativité de chaque axe (basé sur le pourcentage d'inertie). Comme seuls les génotypes complets sont permutés, il est donc important de disposer de jeux de données suffisamment complets si on souhaite que cette procédure ait un minimum de sens.

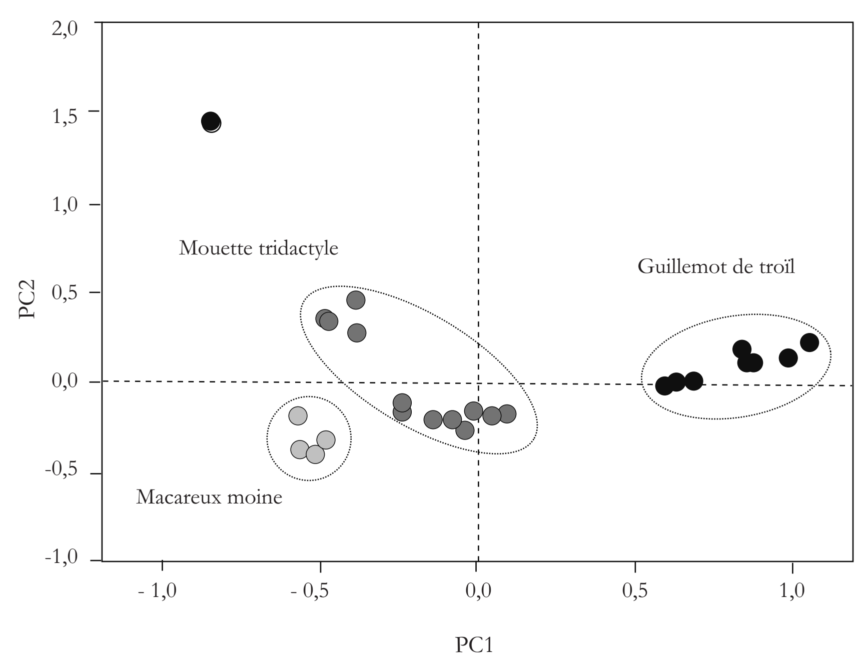


Figure 12
ACP basée sur le polymorphisme de huit loci microsatellites de la tique d'oiseaux coloniaux marins *Ixodes uriae* dans différents sites européens (différents points du graphique) allant des côtes françaises, écossaises et norvégiennes en passant par les îles Faroë. Sur le graphique on voit bien que les différents sous-échantillons se regroupent essentiellement par espèce d'hôtes (points de même couleur) dans le nid desquels les tiques ont été échantillonnées, et non par la localisation géographique, sauf pour les Guillemenots d'Hornøya (Norvège septentrionale) qui se retrouvent excentrés des autres sous-échantillons des tiques de cette espèce (en haut à gauche). Un résultat similaire est également observable dans l'hémisphère sud pour les tiques de différentes espèces de manchots (consulter McCoy *et al.*, 2005). Le pourcentage d'inertie est présenté pour les deux axes, qui se sont montrés significatifs par permutation.

Analyse canonique des correspondances (ACC)

L'ACC (CCA en anglais) est malheureusement implémentée actuellement par un logiciel commercial qui s'appelle CANOCO (Ter Braak, 1986, 1987 ; Ter Braak et Šmilauer, 2002). Il s'agit d'une méthode complexe d'ordination des données visant à directement corréler des tableaux de données multivariées. L'ordination des données couplées aux techniques de régression suivies de tests par permutation des données offre une méthode sophistiquée pour corréler les données génétiques à des variables environnementales. Elle offre également l'opportunité d'obtenir une projection en deux dimensions des centroïdes (barycentres) des données génétiques de chaque sous-échantillon défini, autour de laquelle une ellipse correspondant à l'intervalle de confiance à 95 % de cette projection peut également être dessinée.

L'utilisation de l'ACC est rare, mais peut se montrer efficace ou au moins illustrative (Škalamera *et al.*, 1999 ; Angers *et al.*, 1999). Le logiciel ADE-4 permet aussi ce genre d'approches (Chessel *et al.*, 2004).

Construction d'arbres

Construire des dendrogrammes censés relier les différents individus ou sous-échantillons en fonction de leur proximité génétique procure un moyen assez élégant et relativement simple de représenter les données génétiques suivant un schéma hiérarchique. Ce mode de représentation est d'ailleurs si populaire qu'innombrables sont les études qui l'utilisent. Un des champs d'application privilégié de la construction d'arbres peut être trouvé dans les études d'épidémiologie moléculaire d'organismes clonaux (voir Taylor *et al.*, 1999 pour revue).

Plusieurs méthodes existent. Pour des données type microsatellites ou allozymes, à cause de l'homoplasie, il n'est pas raisonnable d'espérer obtenir quelque chose ayant valeur phylogénétique. Selon Takezaki et Nei (1996), la méthode par NJTREE (*neighbor-joining tree*) basée sur une matrice de distances harmoniques (type Cavalli-Sforza et Edwards, 1967) paraît l'option la plus pertinente. Le logiciel MEGA 5 (Tamura *et al.*, 2011a) (Kumar *et al.*, 2004 ; Tamura *et al.*, 2011b), librement téléchargeable de http://www.megasoftware.net/, offre une interface conviviale pour construire simplement un tel arbre à partir d'une demi-matrice de distances par paire. La méthode UPGMA, qui fait l'hypothèse d'une horloge moléculaire (les branches ont nécessairement la même longueur) est également très souvent utilisée. Je n'ai pas d'opinion bien tranchée sur la question et je doute que l'UPGMA donne souvent des résultats forts différents du NJTREE. Mais comme certains auteurs ont fait des analyses comparatives théoriques ayant conduit à mettre en avant NJTREE et les distances harmoniques, je préfère d'instinct suivre leur recommandation. L'avantage d'utiliser MEGA est que les arbres générés peuvent être sauvés dans le presse-papier (*clipboard*) et collés dans un logiciel de graphique où, après dissociation on peut travailler tranquillement la figure obtenue.

TROUVER UNE SOUS-STRUCTURE CACHÉE

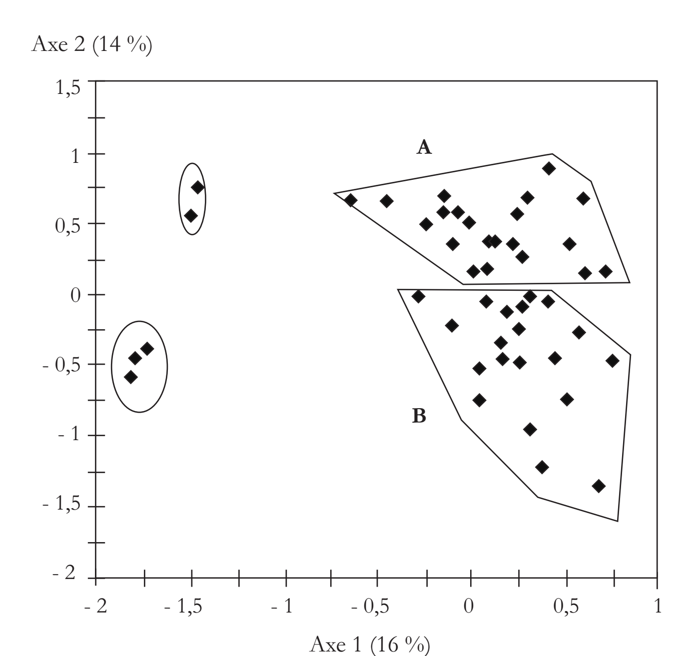
Dans certaines situations, il n'existe aucun indice visible qui permettrait de subdiviser un échantillon en plusieurs sous-unités objectives. Dans de telles situations, les stratégies d'échantillonnage peuvent se montrer inefficaces à représenter une réalité biologique ou écologique existante. En effet, si un facteur biologique et/ou écologique contribue fortement à l'élaboration de l'architecture génotypique des individus

étudiés, on s'attend à ce qu'un tel phénomène laisse une signature génétique visible dans l'échantillon sous la forme d'un déficit en hétérozygotes (effet Wahlund). Le site d'échantillonnage peut, par exemple, correspondre à une aire de nourrissage d'individus provenant d'unités de reproductions très séparées. Il existe différentes méthodes permettant de regrouper les individus d'un échantillon par affinité génétique en différents groupes (sous-populations putatives) en utilisant leur génotype multilocus. Par exemple, de très importants déficits en hérérozygotes sont souvent trouvés pour les loci microsatellites des mouches tsé-tsé. En particulier, Solano *et al.* (2000) sur *Glossina palpalis gambiensis* avaient mis en évidence de très importants F_{IS}. Ces forts déficits ne pouvant être expliqués facilement, et en tous les cas pas en totalité, par la présence d'allèles nuls ou par la dominance d'allèles courts (voir p. 105-111), une structure cachée a été recherchée. À l'époque où ces données ont été analysées, une AFC avait été réalisée et avait permis d'identifier des sous-groupes de tsé-tsé où le déficit en hétérozygotes se retrouvait inférieur au déficit initial (individus regroupés), ce qui confirmait que ces déficits provenaient bien d'un effet Wahlund (dont l'origine exacte reste elle-même à identifier) (fig. 13).

D'autres méthodes, basées sur les statistiques pseudo-bayésiennes et des simulations de Monte-Carlo par chaîne de Markhov, sont maintenant disponibles. Elles permettent d'inférer la vraisemblance avec laquelle certains individus peuvent être regroupés et donc considérés comme appartenant à la même sous-population (*cluster*), ce qui peut conduire à la détection d'une structure cachée. Différentes méthodes avec différents logiciels sont disponibles. On peut citer ici les deux principaux que sont STRUCTURE de Pritchard *et al.* (2002) (Pritchard *et al.*, 2000 ; Falush *et al.*, 2003), librement téléchargeable à http://pritch.bsd.uchicago.edu/software/structure2_1.html, et BAPS 4 de Corander *et al.* (2006) (Corander *et al.*, 2003, 2004 ; téléchargeable à http://www.rni.helsinki.fi/~jic/bapspage.html).

BAPS a notamment permis de détecter une structure cachée dans deux échantillons (séparés dans le temps) de *Glossina palpali palpalis* de Bonon (Côte d'Ivoire) (Ravel *et al.*, 2007).

Dans les deux algorithmes (STRUCTURE et BAPS), l'hypothèse de panmixie est supposée dans chaque cluster que le logiciel cherche à construire. Cependant, la notion de panmixie telle qu'affirmée par les auteurs n'est pas claire et nous verrons que les clusters obtenus ne sont pas nécessairement conformes à Hardy-Weinberg. D'une manière générale, il est très difficile de savoir ce qui est fait et les différences entre ces méthodes mériteraient d'être mieux explorées dans différentes situations de populations structurées (y compris l'absence de structure) et pour différents systèmes de reproduction. Cela signifie qu'il ne faut en aucun cas être esclave du résultat fourni par ces méthodes et garder la tête froide en toute circonstance. D'une manière générale, BAPS est beaucoup plus facile d'utilisation et beaucoup plus rapide, mais produit davantage de clusters que ce qu'il y a en réalité (Latch *et al.*, 2006). INSTRUCT (Gao *et al.*, 2007) fonctionne comme STRUCTURE, mais avec prise



	F_{IS}	
	Échantillon total (A+B)	Échantillons A et B séparés
Locus Gpg553	0,09	0,07
Locus Gpg1962	0,12	0,03
Locus Gpg6922	0,12	0,07
Moyenne sur les loci	0,20	0,03

Figure 13
Résultat de l'AFC sur les génotypes microsatellites des *Glossina palpalis gambiensis* de Nyafaro au Burkina Faso, après retrait d'un certain nombre d'individus trop excentrés (voir p. 96). Les deux grands groupes A et B définis selon l'axe 2 de l'AFC permettent de recalculer le déficit en hétérozygotes (F_{IS}) et de constater une chute entre celui mesuré pour tous les individus regroupés et celui estimé dans les groupes A et B considérés séparément. Les pourcentages d'inertie de chaque axe sont aussi représentés (voir SOLANO *et al.*, 2000 pour plus de détails).

en compte de l'autofécondation. Il ne marche qu'en ligne à http://cbsuapps.tc.cornell.edu/InStruct.aspx, ce qui peut poser un problème pour les pays mal connectés. Enfin, il s'agit d'un domaine en pleine évolution et lorsque ce manuel paraîtra, d'autres logiciels avec d'autres options seront parus. En particulier, dans la seconde partie de ce manuel, nous utiliserons aussi un autre de ces logiciels plus récent,

FLOCK Duchesne et Turgeon (2009), qui bien que différent de BAPS, donne des résultats très comparables et est quasiment aussi facile d'utilisation. Avant analyse, j'engage les lecteurs souhaitant aller plus loin de vérifier sous Google, en recherchant "*admixture AND population structure*", par exemple.

ESTIMER DES EFFECTIFS EFFICACES

Nous avons déjà vu, à propos des tests d'isolement par la distance en p. 89-90, que certains paramètres démographiques sont extrapolables à partir des données génétiques. Il existe d'autres situations où certains paramètres, tels que l'effectif efficace ou le taux de migration, peuvent être inférés à partir de données séparées dans le temps et/ou dans l'espace (sans isolement par la distance).

Définition de l'effectif efficace d'une population

L'effectif efficace d'une population, aussi appelé effectif génétique et généralement noté N_e, est censé représenter avec quelle vitesse une population perd sa diversité génétique par dérive génétique. En effet, la fraction $1/N_e$ ($1/(2N_e)$ pour des diploïdes) donne la probabilité que deux gènes pris au hasard dans la population descendent d'un seul et même gène ancêtre des parents. $1/N_e$ représente aussi la probabilité pour deux gamètes qui s'unissent de provenir du même parent. Ce phénomène est appelé coalescence. Cette coalescence récurrente de certains gènes implique donc que d'autres gènes ne contribuent pas au pool des générations suivantes. Certains gènes sont donc perdus, ce qui signifie également que la diversité génétique s'érode. Le ratio entre la taille réelle de la population N_c (aussi appelée taille de recensement ou *census size* en anglais) et cet effectif efficace représente donc une mesure de la dynamique de la diversité génétique d'une population focale par rapport à une population dite idéale. Une population idéale perdrait sa diversité génétique à la vitesse $1/N_c$ par génération de telle sorte que son effectif efficace serait égal à son effectif de recensement. Une telle situation correspond donc à des populations monoïques à générations non chevauchantes se reproduisant de façon panmictique dans un environnement constant sans sélection, migration ni mutation. À titre d'exemple, une population composée de $N_c = 100$ individus dioïques avec un seul mâle ($N_m = 1$) et 99 femelles ($N_f = 99$) aurait un effectif efficace de (Hartl et Clark, 1989 : 86) :

$$N_e = \frac{4N_m N_f}{N_c} \qquad (58)$$

Ce qui donne un $N_e \approx 4$, soit 25 fois plus petit que la taille de recensement ($N_c = 100$). Ceux qui le souhaitent trouveront une démonstration de ceci en

réponse 10. On comprend aisément qu'une telle population perd sa diversité à une vitesse très grande. D'autres facteurs influencent l'effectif efficace. En particulier, la subdivision des populations est susceptible d'augmenter l'effectif efficace d'une population, car une population subdivisée conservera en général mieux sa diversité génétique qu'une population homogène. Pour caricaturer, une population totalement subdivisée en sous-unités stables n'échangeant aucun migrant aura un effectif efficace infini, car la diversité génétique se trouvera rapidement fixée à l'échelle globale quand chaque sous-population se retrouve fixée pour l'un ou l'autre des allèles présents (et donc quand la diversité est totalement perdue à une échelle locale). Les lecteurs qui n'auront pas encore jeté ce manuel au feu trouveront une excellente revue de Criscione et Blouin (2005) sur le calcul des effectifs efficaces dans le cas des organismes parasites.

Enfin, il faut également signaler qu'il existe plusieurs définitions de l'effectif efficace avec, dans la plupart des situations, des conséquences négligeables sur les valeurs estimées. Citons l'effectif efficace de consanguinité qui, comme son nom l'indique, mesure la probabilité d'identité par descendance des gènes, l'effectif efficace de variance qui s'attache à analyser la variation des fréquences alléliques (leur amplitude plus exactement) d'une génération à l'autre, l'effectif efficace de valeur propre qui s'intéresse à l'évolution de l'hétérozygotie au cours du temps, et enfin l'effectif efficace de coalescence qui s'intéresse au temps qu'il faut mettre pour retrouver l'ancêtre commun de deux représentants du même gène dans une population.

Méthodes de calcul de l'effectif efficace des populations naturelles

Deux familles de méthodes permettent d'inférer l'effectif efficace de populations étudiées, à l'aide de marqueurs moléculaires. Les études basées sur les fluctuations temporelles des fréquences alléliques, au cours des générations, permettent d'estimer ce que les spécialistes appellent l'effectif efficace de variance de populations échantillonnées de façon répétée au cours de leur cycle de vie (Waples, 1989). Le logiciel MACLEEPS 1.1 (Anderson *et al.*, 2000) (téléchargeable à http://www.stat.washington.edu/thompson/Genepi/Mcleeps.shtml) réalise une estimation de N_e par maximum de vraisemblance en utilisant la variation des fréquences des allèles entre générations. Il est donc nécessaire de connaître le temps de génération de l'espèce étudiée. L'algorithme utilisé fait l'hypothèse que la sélection, la migration et la mutation ont un impact négligeable comparé à la dérive. Un intervalle de confiance à 95 % est également calculé (Anderson *et al.*, 2000). Cela est également réalisé par le logiciel NeEstimator (Peel *et al.*, 2004) (voir plus bas).

Les études des variations spatiales des fréquences des allèles permet d'estimer ce que les spécialistes (encore eux) appellent l'effectif efficace de consanguinité. Les estimations produites par ESTIM 1.2 appartiennent à cette catégorie (Vitalis et Couvet,

2001a) (téléchargeable gratuitement par FTP anonyme à ftp://isem.isem.univ-montp2.fr/pub/pc/estim). Ce logiciel utilise l'information monolocus fournie par le F_{ST} et celle offerte par le déséquilibre d'identité par paire de loci afin de pouvoir en tirer N_e et m (effectif efficace et taux de migration) sans avoir besoin de connaître le modèle ni le taux de mutation ou de migration (Vitalis et Couvet, 2001b, c). Cependant, les taux d'autofécondation et de recombinaison entre les loci utilisés doivent eux être connus.

La façon d'utiliser la méthode temporelle (effectif efficace de variance) et la méthode spatiale de Vitalis et Couvet (2001a) (effectif efficace de consanguinité), une comparaison ainsi que les problèmes possibles que l'on peut rencontrer en les utilisant, sont discutés dans Meunier *et al.* (2004b). ESTIM donne souvent des effectifs peu informatifs (0 ou infini). Il fournit également, quand le calcul est possible, les intervalles de confiance à 95 % des valeurs estimées.

Il existe une autre méthode utilisant l'information des déséquilibres de liaison entre loci, dans une seule population (Bartley *et al.*, 1992). Cette méthode, ainsi que celle de Waples (échantillons temporels), est implémentée par le logiciel NeEstimator Version 1.3 (logiciel non publié de Peel D., Ovenden J. R., Peel S. L., 2004, NeEstimator Version 1.3 : *software for estimating effective population size*. Queensland Government, Department of Primary Industries and Fisheries) téléchargeable gratuitement à http://www.dpi.qld.gov.au/fishweb/11629.html. Ce logiciel propose aussi une méthode basée sur les niveaux d'hétérozygotie observés (Luikart et Cornuet, 1999), mais à priori moins précise que la méthode proposée par Balloux (2004) (voir plus bas). Dans tous les cas, et quand le calcul est possible, le logiciel donne les intervalles de confiance à 95 %. La méthode de Bartley étant biaisée quand les échantillons sont de taille inférieure au N_e (England *et al.*, 2006 ; Waples, 2006), Waples et Do proposent LDNe (Waples et Do, 2008) qui donne rarement des résultats utilisables, mais corrige le biais dû aux faibles échantillons et est donc utile pour vérifier qu'on ne sous-estime pas les N_e. Balloux (2004) propose un estimateur corrigé par rapport à la méthode de Luikart et Cornuet et facile à calculer en utilisant l'estimateur de Weir et Cockerham : $N_e = 1/(-2F_{IS}) - F_{IS}/(1 + F_{IS})$.

Enfin, il est important de signaler qu'une approche synthétique, alliant les deux types d'informations (spatiale et temporelle), est également disponible (Wang et Whitlock, 2003). Un logiciel appelé MLNE estimant N_e et m en même temps peut être téléchargé gratuitement à partir de http://www.zoo.cam.ac.uk/ioz/software.htm.

Détection de goulots d'étranglement

Ce paragraphe figure ici car les notions d'effectifs efficaces de goulot d'étranglement (*bottleneck* en anglais) et de biologie de la conservation sont étroitement liées. Une population qui subit une forte réduction d'effectif (goulot d'étranglement) va avoir tendance à présenter une réduction simultanée du nombre d'allèles par locus et de

leur diversité génétique (que nous avons plus haut appelée H_s). Durant un goulot d'étranglement, le nombre d'allèles est réduit plus fortement que la diversité génétique. Il en résulte qu'une population ayant subi un goulot d'étranglement récent présentera une diversité génétique supérieure à celle attendue à l'équilibre mutation/dérive compte tenu du nombre d'allèles observés, sous l'hypothèse d'une taille constante de la population. Plusieurs modèles de mutation peuvent être utilisés selon les situations. Selon Cornuet et Luikart (1996), dans le cas de microsatellites il vaut mieux utiliser le modèle de mutation SMM ou à deux phases, même si c'est avec un IAM que la détection semble la plus aisée. Il s'agit donc de faire un choix entre ce qui, de la détection ou de la non-détection d'un goulot d'étranglement, est plus ou moins grave, ce qui dépend évidemment du contexte. La détection et les tests de significativité de cet excès d'hétérozygotie (signature d'un goulot d'étranglement) sont mis en œuvre dans le logiciel Bottleneck (Piry *et al.*, 1997) (voir Cornuet et Luikart, 1996).

Dans une population à l'équilibre mutation/dérive dont la taille n'a pas varié depuis un temps raisonnable, il y a autant de chance d'observer un excès qu'un déficit de diversité génétique, par rapport à l'attendu, aux différents loci. Afin de détecter si le nombre d'excès observé dépasse significativement ce qui est attendu sous cette hypothèse nulle, on peut utiliser trois tests (décrits par les auteurs dans l'aide du logiciel), mais le plus commode et le plus puissant est le test de Wilcoxon.

Dans leur article, Cornuet et Luikart (1996) montrent (voir leur figure 3) que la détection d'une telle signature ne semble possible que dans certaines conditions, qui dépendent du degré de polymorphisme observé, du nombre de générations écoulées depuis le dernier goulot d'étranglement (qu'on cherche à détecter) et de l'effectif efficace de la population (celui qu'elle acquiert après l'événement de goulot d'étranglement). Par exemple, avec des loci raisonnablement polymorphes (microsatellites), des échantillons inférieurs à 40 individus et moins de 10 loci, la détection d'un goulot d'étranglement n'est possible que si ce dernier a eu lieu dans une fourchette de temps définie par les limites 0,025 x 2 x N_e et 2,5 x 2 x N_e générations et où N_e représente l'effectif efficace qui s'est mis en place après le goulot d'étranglement. Par conséquent, la connaissance de ce temps de générations τ depuis le dernier goulot d'étranglement probable peut offrir une manière détournée d'estimer une fenêtre probable pour N_e. Ici, cette fenêtre serait de $[\tau/5N_e, \tau/0{,}5N_e]$. C'est ce principe qui a permis d'estimer grossièrement les effectifs efficaces probables de la tique du bétail *Rhipicephalus* (*Boophilus*) *microplus* récemment introduite en Nouvelle-Calédonie comme très grands malgré des traitements acaricides soutenus dans les élevages bovins de l'île (Koffi *et al.*, 2006a).

Enfin, il n'est pas inutile de signaler ici que le test de Bottleneck aura tendance à donner des résultats légèrement significatifs quand les populations étudiées sont de petites tailles. Dans ce cas il est utile, voire indispensable, d'obtenir des informations sur la taille des populations étudiées, par l'utilisation de méthodes d'estimation

d'effectifs efficaces, par exemple. Par ailleurs, il ne faudra pencher en faveur d'un goulot d'étranglement que si les *P*-values sont très significatives et/ou si au moins deux, et encore mieux les trois, modèles de mutation convergent vers cette conclusion.

L'utilisation de plusieurs méthodes pour estimer N_e pourra être d'un grand secours pour convaincre les referees toujours réticents (si les valeurs obtenues par différentes méthodes convergent, bien entendu).

LE CAS SPÉCIAL DES ALLÈLES NULS

Présentation générale

Les allèles nuls correspondent à des allèles qu'on ne peut pas déceler avec la méthode de détection biochimique utilisée. Ils sont invisibles à l'état hétérozygote, car récessifs par rapport aux autres allèles, et mal détectés à l'état homozygotes (blancs), car il est souvent difficile de séparer les cas où la manipulation a échoué (mauvaise amplification, matériel dégradé, etc.) des cas où on a réellement à faire à un homozygote nul. Les allèles nuls sont fréquemment rencontrés dans les études de génétique des populations naturelles, bien que fréquemment ignorés. Il est même probable que bon nombre de déficits en hétérozygotes documentés dans de nombreux articles soient en fait dus à ce phénomène, alors que d'autres causes sont privilégiées dans les articles en question. Les allèles nuls peuvent être fréquents même dans le cas des allozymes (Gaffney, 1994 ; Nébavi *et al.*, 2006), où on ne les attend pourtant guère, car ils correspondent dans cette circonstance à des enzymes non fonctionnels, bien qu'indispensables à la vie (pour la plupart). On augurerait donc ici une moindre valeur sélective des allèles nuls, à moins qu'un mécanisme permette de les garder à l'état hétérozygote le plus fréquemment possible, comme cela peut être le cas chez les organismes clonaux (Nébavi *et al.*, 2006). C'est un problème rencontré typiquement chez les marqueurs microsatellites (Paetkau et Strobeck, 1995 ; Pemberton *et al.*, 1995 ; Brookfield, 1996). Une mutation dans la séquence flanquante, au niveau des séquences correspondant à un des *primers*, empêche la bonne amplification de cet allèle. Il apparaîtra « blanc » (aucun signal) à l'état homozygote et sera dominé par les allèles avec lesquels il sera hétérozygote. Les hétérozygotes pour ce type d'allèles apparaissent donc homozygotes pour l'autre allèle.

Détecter la présence d'allèles nuls

Nous savons maintenant que la présence d'allèles nuls à un locus va provoquer des déficits en hétérozygotes inexplicables biologiquement. Normalement, dans ce cas, on s'attend à ce que les différents loci donnent une mesure différente (variance

entre loci) (De Meeûs *et al.*, 2002a ; Hurtrez-Boussès *et al.*, 2004). Par ailleurs, s'il y a structuration entre sous-échantillons, la fréquence de ces allèles nuls, aux loci concernés, devrait changer d'un sous-échantillon à l'autre et provoquer une variance des déficits (F_{IS}) entre sous-échantillons, mais seulement pour des niveaux de différenciation génétique élevés. Ensuite, il existe des procédures plus ou moins complexes pour estimer, à chaque locus et dans chaque sous-échantillon, la fréquence d'allèles nuls nécessaires pour expliquer les déficits observés (Brookfield, 1996). Le logiciel Micro-checker V 2.2.3. (Van Oosterhout *et al.*, 2004), téléchargeable librement de http://www.microchecker.hull.ac.uk/), permet de faire ces estimations pour chaque locus et chaque sous-échantillon. Ces fréquences estimées d'allèles nuls peuvent ensuite permettre d'évaluer la proportion attendue d'individus blancs, sous l'hypothèse de panmixie et si ces allèles nuls expliquent la totalité du déficit.

Trucs et astuces pour tester la présence des allèles nuls

Il est important d'insister encore sur le fait que, si tous les loci convergent vers le même déficit en hétérozygotes (tous présentent un F_{IS} comparable à celui des autres), il n'est alors pas nécessaire d'invoquer les allèles nuls, mais plus parcimonieusement une cause biologique (autofécondation, effet Wahlund). Dans ce qui suit, nous partons donc du principe qu'une forte variance entre loci a été observée.

Il faut tout d'abord savoir que Micro-checker est conçu spécifiquement pour les microsatellites. Avant de procéder à l'analyse, il est demandé le type de motif pour chaque marqueur. Si vous n'êtes pas sûr de vos données (quelques mutants atypiques d'un pas différent du motif de base), il vaut en général mieux adopter l'option mononucléotidique pour tous les loci. Ensuite, parmi les résultats que propose Micro-Checker, il faut garder, pour chaque locus et chaque sous-échantillon, la fréquence des allèles nuls, ainsi que la présence ou non de stuttering. Il vaut mieux utiliser la méthode 2 de Brookfield (1996) qui tient compte des données manquantes (doubles nuls). La fréquence attendue d'homozygotes blancs sera, sous l'hypothèse de croisement au hasard, égale à p_{nul}^2. On peut comparer par un test binomial cette fréquence attendue aux nombres de blancs effectivement observés à ce locus dans le sous-échantillon concerné. Ce test permet de vérifier si les allèles nuls expliquent raisonnablement les déficits observés aux loci concernés. On peut demander à Michel Raymond (Isem, université Montpellier-2) le droit d'utiliser un petit programme qu'il avait conçu et qui s'appelle Multinom, effectuant tous les tests multinomiaux possibles, ainsi que, puisque « qui peut le plus peut le moins », le test binomial. On peut également utiliser le logiciel R qui est gratuit avec la commande binom.test (solution réellement la plus simple).

MicroChecker ne peut pas travailler avec des sous-échantillons trop petits (il renvoie un message d'erreur dans ce cas). Si la plupart de vos sous-échantillons

renvoient un message d'erreur dû à la faiblesse des effectifs, il existe encore une alternative. Vous pouvez effectuer la régression du F_{IS} observé par locus et sous-échantillon en fonction du nombre de blancs observés par locus et sous-échantillon. Si la régression est significative, c'est que les allèles nuls expliquent une partie du F_{IS}. Cette méthode, en calculant le R^2 de la régression (proportion de la variance expliquée par la régression), permet aussi d'appréhender à quel point les allèles nuls expliquent les données.

Toutes ces méthodes font l'hypothèse qu'il y a *grosso modo* panmixie. Si la fréquence de nuls ne permet pas d'expliquer correctement tous vos déficits en hétérozygotes, en particulier si aucun locus n'est conforme à l'attendu panmictique, il se peut que d'autres phénomènes soient en cause. Si par exemple, il y a de l'autofécondation ou effet Wahlund, en plus des allèles nuls, nous ne pourrons pas expliquer les données à l'aide des seuls allèles nuls. Dans le cas de l'autofécondation, il existe un logiciel permettant d'estimer le taux d'autofécondation en tenant compte des allèles nuls ou autre problème (dominance partielle, dominance des allèles courts). Il s'agit de RMES (David *et al*, 2007), qui fait l'hypothèse d'équilibre de liaison entre loci et utilise les déséquilibres d'hétérozygotie par paire de loci, ce qui peut poser un problème dans les petites populations très autofécondantes. Le logiciel est librement téléchargeable à http://www.cefe.cnrs.fr/genetique-et-ecologie-evolutive/patrice-david.

LE CAS TRÈS SPÉCIAL DE LA DOMINANCE DES ALLÈLES COURTS

Point de vue théorique

La dominance des allèles courts, ou « *short allele dominance* » ou encore « *large allele dropout* », est un phénomène rare, mais possible (Wattier *et al.*, 1998 ; De Meeûs *et al.*, 2004a). La logique qui se cache derrière ce terme est la suivante. Si, par un mécanisme qui reste à mettre en évidence, une compétition existe, au cours de la PCR, entre les deux portions d'ADN correspondant aux deux allèles d'un même locus devant être amplifiés, alors il semble logique que ce soit l'allèle le plus court (s'il y a une différence de taille entre les deux, bien entendu) qui sera le mieux amplifié.

Une tentative de modélisation du phénomène peut être trouvée dans De Meeûs *et al.* (2004a). Dans ce modèle, on suppose une population panmictique et un locus pour lequel la PCR favorisera l'amplification de l'allèle le plus court de façon proportionnelle à la différence de taille qui l'oppose à l'autre allèle, ainsi qu'à un paramètre α variant entre 0 (pas de dominance) et 1 (dominance totale des allèles courts). Si les allèles existant à ce locus se rangent dans un ordre de tailles croissantes

de s_1 à s_n et que la fréquence d'un allèle quelconque i est notée p_i, on peut poser que la fréquence observée d'hétérozygotes pour cet allèle avec un autre allèle j sera de :

$$2p_i p_j \left[1-\alpha \frac{s_i - s_j}{s_n - s_1}\right] \text{ si } s_i > s_j \tag{59}$$

et

$$2p_i p_j \left[1-\alpha \frac{s_j - s_i}{s_n - s_1}\right] \text{ si } s_i < s_j \tag{60}$$

Dans les équations (59) et (60), on voit bien que le biais sera maximal pour l'écart maximal de taille, c'est-à-dire pour un hétérozygote pour les allèles 1 et n, et minimal entre deux allèles les plus proches. En utilisant ces deux équations, on en déduit que la proportion observée d'hétérozygotes pour l'allèle i avec tous les autres allèles sera de :

$$H_i = \sum_{j=1}^{j=i-1} 2p_i p_j \left[1-\alpha \frac{s_i - s_j}{s_n - s_1}\right] + \sum_{j=i+1}^{n} 2p_i p_j \left[1-\alpha \frac{s_j - s_i}{s_n - s_1}\right] \tag{61}$$

soit :

$$H_i = 2p_i \left\{(1-p_i) - \frac{\alpha}{s_n - s_1}\left[\sum_{j=1}^{j=i-1}(s_i - s_j)p_j + \sum_{j=i+1}^{j=n}(s_j - s_i)p_j\right]\right\} \tag{62}$$

L'équation (63) nous donne donc l'hétérozygotie attendue sous panmixie moins la proportion des génotypes hétérozygotes erronément interprétés comme homozygotes pour le plus court des allèles. La proportion d'homozygotes observés pour l'allèle i sera donc celle attendue sous panmixie plus la proportion d'individus portant l'allèle i et un allèle plus long que j et interprétés comme homozygotes pour cet allèle. Cette homozygotie observée sera donc de :

$$F_i = p_i \left\{p_i + 2\frac{\alpha}{s_n - s_1}\sum_{j=i+1}^{j=n}(s_j - s_i)p_j\right\} \tag{63}$$

L'utilisation des équations (1), (62) et (63) nous permet alors d'estimer la fréquence erronément observée de l'allèle i dans l'échantillon comme :

$$p_i' = \frac{2F_i + H_i}{2} = F_i + \frac{1}{2}H_i \tag{64}$$

En utilisation l'équation (6), nous pouvons alors estimer le déficit artificiel d'hétérozygotes observé par rapport aux attendus panmictiques :

$$F_{IS_i} = 1 - \frac{H_i}{2p_i'(1-p_i')} \tag{65}$$

Dans la figure 14, il n'est pas inintéressant de constater que la relation entre taille des allèles et leur déficit en hétérozygotes n'est ni linéaire ni monotone et dépend de

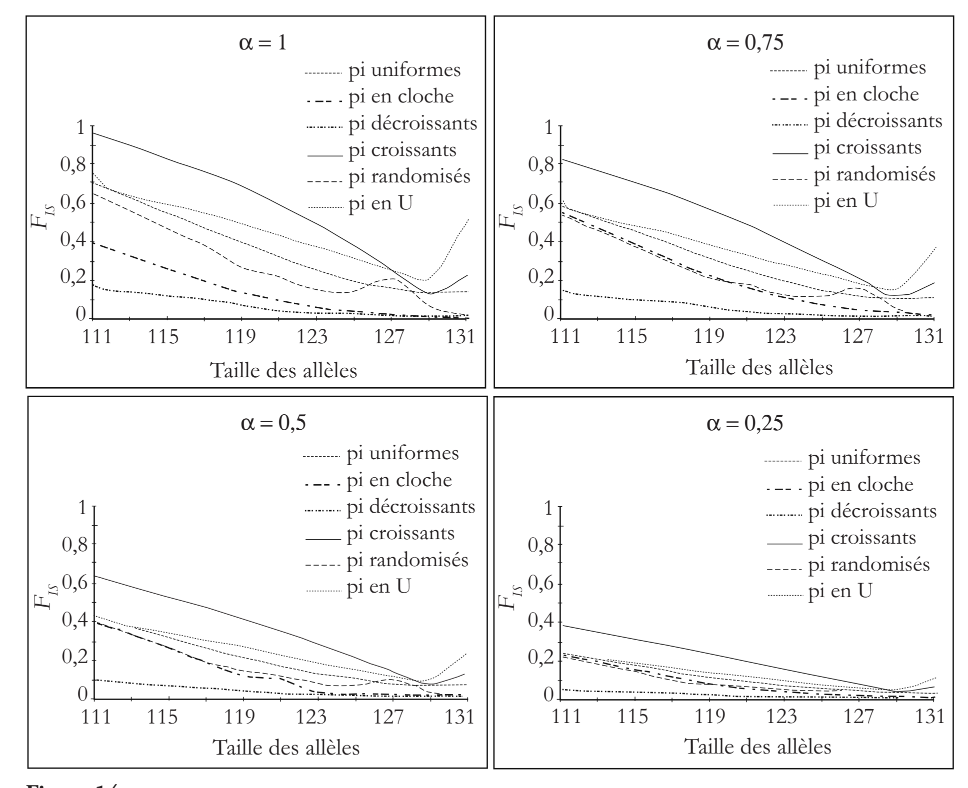


Figure 14
Évolution du F_{IS} en fonction de la taille des allèles pour une gamme de taille microsatellites allant de 111 à 131 paires de bases, pour différentes distributions de fréquences des allèles (voir le tableau 5) et pour différentes valeurs de dominance des allèles courts (α).

la distribution des fréquences des allèles (le tableau 5 décrit les différentes distributions utilisées), mais globalement on s'attend quand même à observer une décroissance de F_{IS} en fonction de la taille des allèles.

On constate aussi que ce phénomène modifie également l'estimation des fréquences des allèles.

Du point de vue pratique : détection de la dominance des allèles courts

Il existe une procédure de détection de la dominance des allèles courts dans le logiciel Micro-Checker, appelée ici « *large allele drop-out* ». Mais cette procédure ne teste le phénomène que dans chaque sous-échantillon pris séparément. Il en résulte un manque de puissance. On peut tester l'existence d'une dominance d'allèles courts sur l'ensemble des sous-échantillons en utilisant une approche de régression.

Tableau 5
Distributions de fréquences des allèles utilisées pour étudier la relation entre F_{IS} et taille des allèles dans le cadre d'une dominance des allèles les plus courts (voir la figure 14).

Allèles	Uniformes	En cloche	Décroissants	Croissants	Randomisés	En U
111	0,1000	0,0250	0,6000	0,0025	0,0100	0,3000
112	0,1000	0,0550	0,2000	0,0025	0,0025	0,1000
117	0,1000	0,1000	0,1000	0,0050	0,6000	0,0600
119	0,1000	0,1600	0,0500	0,0100	0,0025	0,0300
121	0,1000	0,3000	0,0200	0,0100	0,0200	0,0100
123	0,1000	0,1700	0,0100	0,0200	0,0050	0,0100
125	0,1000	0,1000	0,0100	0,0500	0,1000	0,0300
127	0,1000	0,0500	0,0050	0,1000	0,2000	0,0600
129	0,1000	0,0270	0,0025	0,2000	0,0500	0,1000
131	0,1000	0,0130	0,0025	0,6000	0,0100	0,3000

Tout d'abord, pour le locus étudié, il faut récupérer le F_{IS} de chaque allèle dans chaque sous-échantillon. Fstat ne le fait malheureusement pas automatiquement. Il faut créer autant de fichiers Fstats qu'il y a de sous-échantillons et, dans chacun de ces fichiers, il faut créer une population fictive fixée (un seul allèle présent) pour les locus dont on veut les F_{IS} par allèle. Fstat n'aime en effet pas travailler sur une seule population. Une fois qu'on a fait calculer ces F_{IS} par le logiciel, on a tout ce qui est nécessaire pour effectuer une régression linéaire généralisée ou GLiM. GLiM (*Generalised Linear Model*) est une forme de régression qui permet d'analyser des données de n'importe quelle forme (gaussiennes, poissoniennes, logistiques pour les plus utilisées) en fonction de n'importe quel type de variable (facteur catégoriel, logique, ordinal discontinu ou continu). Cette régression doit donc être de la forme $F_{IS} = S + T + \text{Cte}$, avec S pour le sous-échantillon, T la taille de l'allèle et Cte une constante. Une analyse de variance sur le modèle permet ensuite de tester si l'effet de la taille des allèles, corrigé de l'effet des sous-échantillons, est significatif ou non. Attention, il faut que la relation entre taille des allèles et F_{IS} soit négative. Les relations positives doivent donc être ignorées.

Nous verrons tout ceci en détail dans la mise en pratique de toutes ces connaissances dans la deuxième partie.

LE CAS DU « *STUTTERING* »

Le mot anglais « *stuttering* » se traduit par bégaiement. Si les amorces de la PCR ne s'accrochent pas très bien, il est possible alors que la Taq polymérase bégaie et amplifie deux fois un al lèle de même taille pour un individu hétérozygote si les deux allèles ont une taille proche. Il doit donc exister une relation entre déficit observé et différence de taille entre allèles, le déficit devant être plus important pour les individus hétérozygotes pour des allèles de tailles proches. Ceci est détecté automatiquement par Micro-Checker.

Applications à des exemples concrets

Il n'est pas nécessaire de préciser qu'avoir lu la première partie de ce manuel avant d'attaquer la partie pratique facilitera grandement la lecture et la compréhension de cette section, même si on peut très bien commencer directement ici. Je considérerai les notions de génétique des populations et de statistiques utilisées comme un minimum connues. Je ne m'étendrai donc jamais sur un concept ou une notion. Dans le doute, les lecteurs sont invités à se référer aux chapitres de la partie précédente de ce manuel.

Tous les jeux de données utilisés dans cette partie sont disponibles sur internet, à télécharger sur mon site web à http://gemi.mpl.ird.fr/SiteSGASS/SiteTDM/Data/. Tous les logiciels utilisés ou presque sont gratuits. En ce qui me concerne, j'utilise Excel (Microsoft corporation) pour gérer mes données, faire des calculs (transformations de données, par exemple) et des graphiques (comme des courbes). Pour les analyses statistiques classiques, j'utilise des logiciels commerciaux dont j'ai la licence. Cependant, dans un souci de libre accès à tous, j'ai essayé d'adapter tous les tests utilisés pour des logiciels gratuits (voir la liste des logiciels et URL de téléchargement en annexe).

Tous les jeux de données analysés ont fait l'objet d'articles publiés dans des revues scientifiques. Cependant, toutes les analyses présentées dans ce manuel n'ont pas été publiées pour des contraintes d'espace et de lisibilité des articles. On ne publie en général pas les simulations et/ou analyses annexes redondantes que l'on peut être amené à faire pour vérifier la robustesse de certains résultats. Certaines améliorations, comme l'utilisation d'une méthode plus puissante non disponible à l'époque de l'article, ou parce que je n'y avais simplement pas pensé à l'époque, sont également présentées dans certains traitements des données et donc certaines conclusions peuvent parfois être quelque peu modifiées par rapport à l'article princeps.

4 La tique *Ixodes ricinus* et les pathogènes (*Borrelia* sp.) qu'elle transmet

INTRODUCTION



Ce jeu de données, publié dans trois articles (De Meeûs *et al.*, 2002a, 2004a, 2004b), représente un excellent exercice, car nous allons y rencontrer bon nombre de situations décrites dans le chapitre précédent. Nous allons entièrement décortiquer une nouvelle fois ce jeu de données avec les mêmes méthodes, mais aussi avec des outils plus récents que ceux qui avaient été utilisés à l'époque, ce qui sera aussi intéressant. Nous repartirons de zéro en feignant d'ignorer ce qui a déjà été fait, comme s'il s'agissait d'un jeu de données non analysé. Le jeu de données complet est téléchargeable sur mon site web.

ÉTAT DES LIEUX

Les tiques sont des acariens hématophages qui, au cours de leur repas sanguin, peuvent transmettre des maladies à leurs hôtes vertébrés. Dans l'hémisphère nord, ce sont elles qui sont responsables de la très grande majorité des maladies à vecteur des humains et, en particulier, de la transmission de la maladie de Lyme dont l'impact économique et en santé publique est reconnu (Gubler, 1998). Encore aujourd'hui, beaucoup reste à faire pour mieux comprendre l'épidémiologie de cette maladie et la variabilité des manifestations cliniques qui la caractérise (Hubbard *et al.*, 1998). Les tiques sont typiquement des organismes difficiles à suivre sur le terrain, et des approches par marqueur moléculaire semblent donc pertinentes dans ce cas de figure. Après une tentative peu fructueuse avec les allozymes, avec seulement deux loci peu polymorphes (Delaye *et al.*, 1997), des microsatellites ont été développés (Delaye *et al.*, 1998). Seuls cinq loci polymorphes avaient pu être mis au point à l'époque, ce qui était vraiment peu. Nous allons ensemble voir que, malgré cela et les problèmes rencontrés, on peut quand même recueillir beaucoup d'informations pertinentes à l'aide des méthodes décrites dans ce manuel.

En téléchargeant le fichier "IRTotBrut.txt", vous aurez les données brutes obtenues sur des tiques adultes échantillonnées sur la végétation (donc non gorgées), sauf pour

la Tunisie où les tiques étaient fixées sur des vaches. Le fichier "IRTotBrut.txt" est un fichier texte mais que l'on peut ouvrir sous Excel si on le souhaite. Le tableau 6 donne un extrait du fichier de données brutes. Le fichier comprend neuf colonnes. La première colonne donne le nom des sites où les tiques ont été échantillonnées. Il y a huit sites en Suisse (fig. 15) et un site en Tunisie. La deuxième colonne correspond à l'année d'échantillonnage, car certains sites ont été prélevés aux printemps 1995 et 1996 et d'autres uniquement au printemps 1996. La troisième colonne correspond au sexe de la tique (F pour femelle et M pour mâle). La quatrième colonne donne le nom codé des différents individus tiques. Ce codage individuel peut être utile si on fait des analyses individus centrées telle qu'une AFC ou une construction d'arbre sur distances interindividuelles. Enfin, les cinq dernières colonnes correspondent aux génotypes (en taille d'allèles) aux cinq loci microsatellites polymorphes définis dans Delaye *et al.* (1998).

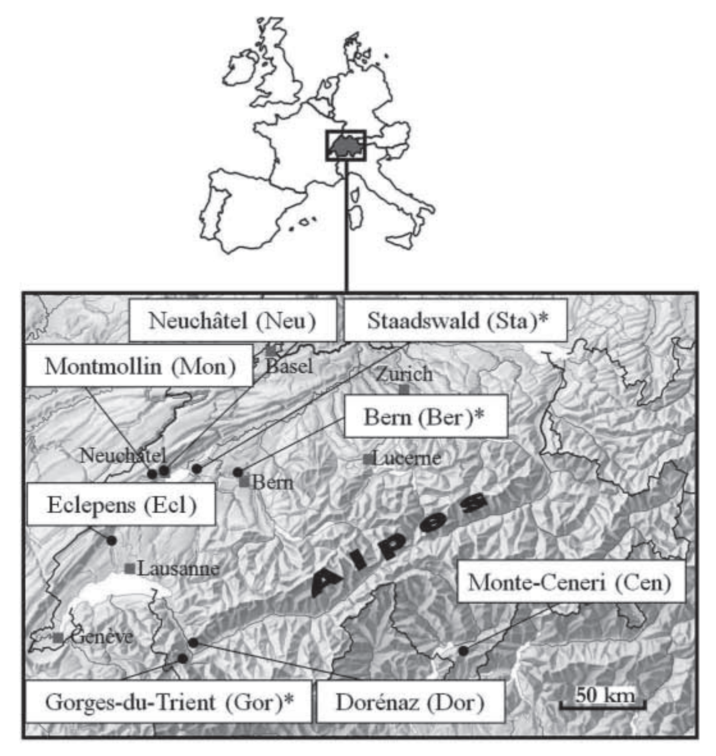


Figure 15
Localisation des sites d'échantillonnage des tiques *Ixodes ricinus* en Suisse et abréviations du nom des sites. Les sites marqués avec un astérisque ont été échantillonnés en 1995 et 1996.

Tableau 6
Extrait du fichier de données IRTotBrut.txt.

Site	Année	Sexe	Individu	IR08	IR25	IR27	IR32	IR39
Bern	95	F	Bern95F_005	170183	150150	123123	235235	129129
Bern	95	F	Bern95F_007	174174	137146	119119	233250	133133
Bern	95	F	Bern95F_011	177183	000000	119119	243243	000000
Bern	95	F	Bern95F_013	173175	136142	119119	250250	142142
Bern	95	F	Bern95F_018	165178	137146	119119	243248	142142
Bern	95	F	Bern95F_020	165173	145148	119119	241241	129133
Bern	95	F	Bern95F_022	168171	134134	119119	243248	135135
Bern	95	F	Bern95F_027	171175	147147	119119	233233	125125
Bern	95	F	Bern95F_028	169175	140145	119119	233233	135142
Bern	95	F	Bern95F_029	166176	128145	119119	243243	125142
Bern	95	F	Bern95F_032	173183	134134	121121	233233	131137
Bern	95	F	Bern95F_037	175183	147147	119119	235235	134137
Bern	95	F	Bern95F_038	175183	135147	123123	250250	127127
Bern	95	F	Bern95F_039	183183	134134	119119	233243	121128
Bern	95	F	Bern95F_040	168174	141147	119119	233233	135142
Bern	95	F	Bern95F_042	174178	146146	119119	000000	112129
Bern	95	F	Bern95F_043	175175	000000	123123	233235	127134
Bern	95	F	Bern95F_044	174176	130130	119119	233233	128128
Bern	95	F	Bern95F_045	171175	145145	119121	243246	142142
Bern	95	F	Bern95F_048	173183	147147	119119	243243	129142
Bern	95	F	Bern95F_049	168170	000000	119121	233233	131144
Bern	95	F	Bern95F_050	169169	150151	119119	233233	129135
Bern	95	M	Bern95M_006	177177	134147	119119	233233	129129
Bern	95	M	Bern95M_008	172172	137148	119119	000000	000000
etc.								

PREMIER RECODAGE DES DONNÉES

Certains logiciels n'aiment pas les noms longs et encore moins les accents ou autres signes cabalistiques. Par ailleurs, il est plus commode pour la lisibilité que tous les noms d'un même niveau aient le même nombre de caractères (alignement des colonnes). C'est pourquoi j'ai choisi de recoder dans IRTotBrut1.txt le nom des sites qui a été raccourci. Dans les données initiales, certains individus sont apparus avec trois ou quatre bandes à certains loci. Nous avions codé ces génotypes 333000 et 444000 pour les génotypes à trois et quatre bandes respectivement. Il convient de recoder ces données en données manquantes (000000). Nous reviendrons sur ces génotypes bizarres un peu plus tard, car ils s'avéreront utiles pour discuter des résultats des analyses de pedigrees. Nous allons procéder à une première analyse avec tous les échantillons afin de tester la panmixie locale et les déséquilibres de liaison entre loci. Nous allons pour ce faire créer un nouveau fichier où les sites et les dates seront distingués, mais aussi le sexe des tiques car on ne sait jamais à l'avance si des différences peuvent exister entre les deux sexes (PRUGNOLLE et DE MEEÛS, 2002 ; PRUGNOLLE *et al.*, 2003), auquel cas les résultats obtenus pourraient s'en ressentir, mais surtout la discussion serait réorientée. Donc autant distinguer le sexe des individus dès le départ, quitte à ignorer ce facteur par la suite si on ne voit rien. Nous allons nommer ce fichier "IRTotTestPanmix.dat" et le mettre au format Fstat qu'il faut donc télécharger et ouvrir pour voir comment constituer un fichier à ce format. Vous pourrez aussi créer un fichier contenant le nom des sous-échantillons "IRTotTestPanmix.lab", car un fichier de données Fstat ne contient que des chiffres. Ce fichier est constitué d'une colonne avec le nom des sous-échantillons. Vous pourrez aussi coder les données au format CREATE (qui n'existait pas au moment de réanalyser ces données) et vous servir de ce logiciel pour convertir ce fichier au format approprié.

PREMIÈRES ANALYSES : INDÉPENDANCE ENTRE ALLÈLES DANS ET ENTRE LOCI DANS LES SOUS-ÉCHANTILLONS

Nous allons donc tester s'il existe des déficits en hétérozygotes et des déséquilibres de liaison. Pour ce faire, il faut ouvrir Fstat. Une fois dans Fstat, il faut ouvrir le fichier "IRTotTestPanmix.dat" et cocher les cases qui vont nous être utiles ici (fig. 16). Si vous souhaitez voir apparaître les noms des sous-échantillons, il faut le

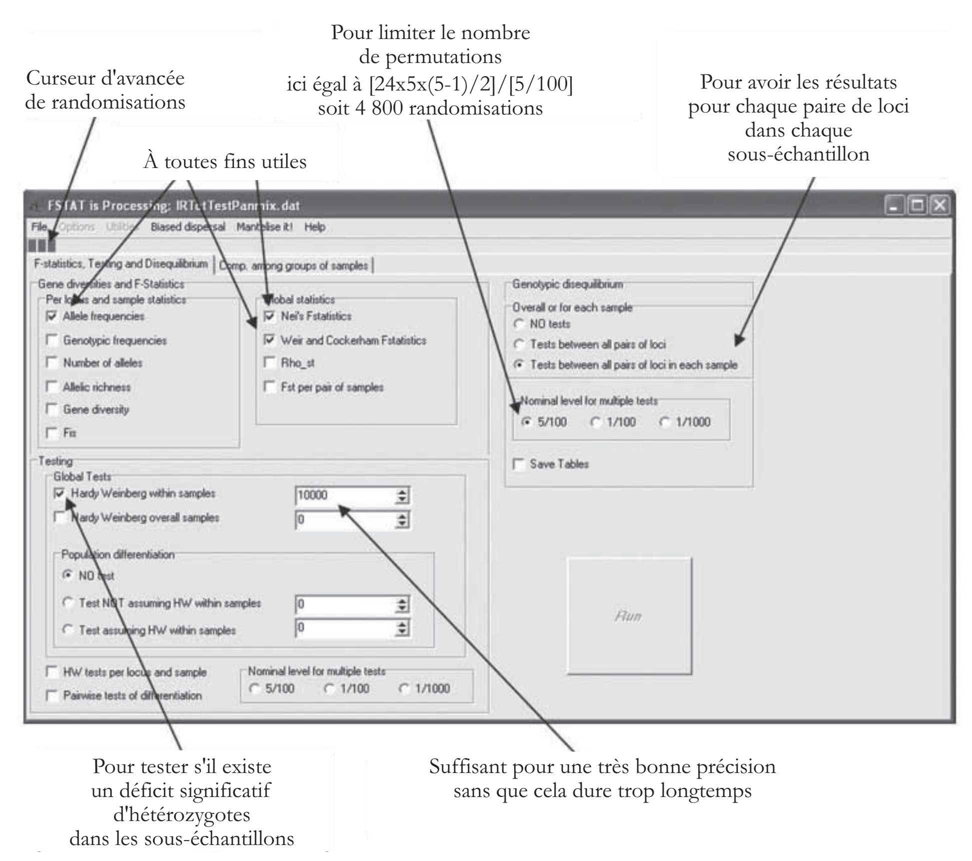


Figure 16
Capture d'écran de Fstat lors de la première analyse.

spécifier par le menu “Options” de Fstat (cf. le premier recodage des données du chapitre 2 de cette deuxième partie pour une prise en main pas à pas de Create).

Nous n'effectuons pas d'autres analyses pour le moment, car ces dernières pourraient être remises en cause par les résultats obtenus ici.

La procédure de test de déséquilibre de liaison est assez lente, donc, si vous souhaitez que votre analyse finisse avant l'âge de la retraite, il vaut mieux dans tous les cas s'en tenir à l'option 5/100 pour le “Nominal level for multiple testing”. Mon ordinateur portable, dont l'horloge à 2.13 GHz et la mémoire vive à 2 Go témoignent d'une performance somme toute raisonnable, a mis quand même quatre heures pour effectuer cette première analyse dont le résultat est consultable dans le fichier “IRTotTestPanmix.out”. Que pouvons-nous voir dans ce fichier ?

Les premières lignes donnent les fréquences des allèles pour chaque locus et chaque sous-échantillon, ainsi que sur l'ensemble (moyennes pondérée, W, et non pondérée,

UW). Nous pouvons constater à cette occasion que chaque locus, sauf IR27, possède un très grand nombre d'allèles dont la plupart ne suivent en rien le modèle de mutation attendu de deux pas par deux pas (ce sont tous des dinucléotides). Dans ce cas, la plupart des allèles proviennent de mutations intervenues en dehors du motif microsatellite, dans les séquences flanquantes. Ce n'est pas dramatique même si non idéal. Suivent les estimateurs de Nei, en particulier ceux des diversités géniques intra-sous-échantillons (H_s) et globale (H_T). Ensuite, les résultats des tests de déséquilibre de liaison sont donnés par paire de loci et par sous-échantillon et sur l'ensemble des sous-échantillons (mais toujours par paire de loci). La mention "*Adjusted P-value for 5 % nominal level is : 0,000208*" ne doit pas vous inquiéter. C'est le calcul du seuil de Bonferroni sur l'ensemble des tests réalisés. Comme il y a 24 sous-échantillons, cinq loci et donc 5(5 - 1)/2 paires de loci, cela correspond à 240 tests. Le seuil corrigé par la procédure de Bonferroni à α = 0,05 est donc α' = 0,05/240 = 0,000208, seuil rarement (jamais ?) accessible, ce qui illustre une discussion que nous avons déjà eue précédemment. De toutes manières, nous ne regarderons ici que les tests multi-sous-échantillons (colonne "All") et donc au pire, le seuil est à diviser par 10, ce qui est inutile puisque nous pouvons aussi constater qu'aucun déséquilibre de liaison n'est significatif. Les loci sont donc raisonnablement indépendants statistiquement les uns des autres. Nous pouvons donc sereinement oublier ces derniers et passer à la suite.

Suivent les estimateurs de Weir et Cockerham dont un seul nous intéresse pour le moment, *f*, l'estimateur du F_{IS}, par locus, par allèle et sur l'ensemble des allèles, sur l'ensemble des loci. Puis suivent les résultats des jackknives et bootstraps et enfin des permutations. En compilant ces résultats dans le tableau 7 et la figure 17, nous constatons de très forts et très variables déficits en hétérozygotes (tous très significatifs avec des *P*-values toutes inférieures à 0,0001, visibles en fin de fichier).

Tableau 7
Valeurs moyennes de *f*, estimateur du F_{IS}, par locus et intervalle de confiance tels que définis par Li et Ls (limite inférieure et supérieure) obtenus pour les microsatellites d'*Ixodes ricinus*. Pour chaque locus, Li et Ls sont calculées à l'aide de l'erreur standard (StdErrFis) donnée par le jackknife sur les populations et la valeur du *t* pour 23 ddl (24 - 1) et α = 0,05 (soit 2,069, voir le tableau 3) en suivant l'équation (45). Pour la valeur globale, l'intervalle de confiance est issu du bootstrap sur les loci.

	IR08	IR25	IR27	IR32	IR39	Global
Moyenne	0,489	0,490	0,490	0,624	0,315	0,475
Li	0,286	0,440	0,422	0,533	0,253	0,386
Ls	0,692	0,540	0,558	0,715	0,377	0,562
StdErrFis	0,098	0,024	0,033	0,044	0,03	

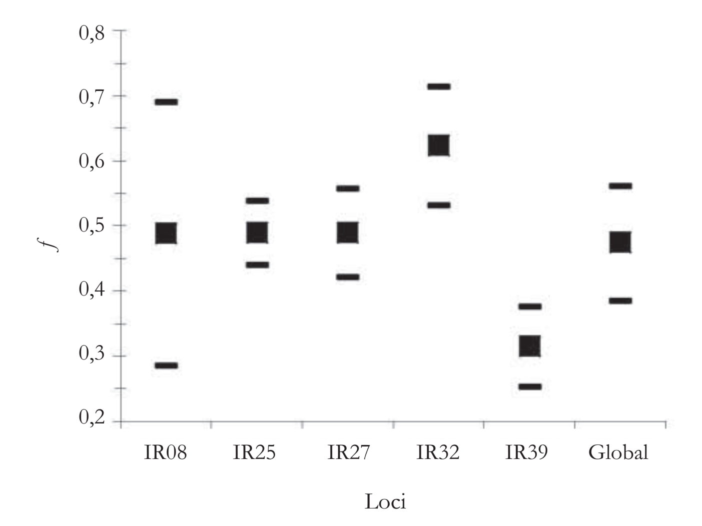


Figure 17
Valeurs moyennes de *f*, estimateur du F_{IS}, par locus et intervalle de confiance obtenus pour les microsatellites d'*Ixodes ricinus*. Pour chaque locus, les intervalles de confiance sont calculés à l'aide de l'erreur standard donnée par le jackknife sur les populations et la valeur du *t* pour 23 ddl (24 – 1) et α = 0,05 (soit 2,069, voir le tableau 3) en suivant l'équation (45). Pour la valeur globale, l'intervalle de confiance est issu du bootstrap sur les loci.

Ces fortes valeurs sont aberrantes étant donné qu'on sait qu'*I. ricinus* pratique une reproduction bi-parentale obligatoire. Des croisements systématiques entre apparentés pourraient-ils expliquer un F_{IS} = 0,5 ? Dans la réponse 11, on décrit comment obtenir une estimation grossière du taux de croisements frère-sœur *b* nécessaires pour expliquer un F_{IS} donné :

$$b = \frac{4F_{IS}}{1+3F_{IS}} \tag{66}$$

Par conséquent, nous avons besoin ici de 4/5, soit 80 % de croisements frère-sœur pour expliquer nos données, ce qui est possible mais semble peu réaliste. *Ixodes ricinus* est en effet une tique triphasique qui change d'hôte pour chaque stade. Les adultes dont nous analysons la variabilité génétique ont donc subi deux phases de dispersion par des hôtes différents. Pour permettre un taux de 80 % de croisements frère-sœur, il faut admettre que 80 % des individus d'une même ponte restent ensemble au cours des différents stades (larvaire, nymphal et adulte) de leur vie.

Il se pourrait, contrairement à ce qui est observé en laboratoire où aucun œuf non fécondé n'a pu éclore, que cette espèce pratique une parthénogénèse automictique d'un type qui augmente l'homozygotie (pour des descriptions des différents modes

d'automixie, voir par exemple De Meeûs *et al.*, 2007b). Seules les femelles sont en général capables de parthénogénèse. Il existe cependant une espèce de cyprès et une espèce de phasme où les mâles se reproduisent asexuellement (voir encore De Meeûs *et al.*, 2007b) et une espèce de fourmi où mâles et femelles sont clonaux chacun de leur côté (Foucaud *et al.*, 2010). Mais ce sont des exceptions. Si parthénogenèse il y a, les femelles devraient donc présenter de beaucoup plus gros déficits en hétérozygotes que les mâles (tous issus d'une reproduction croisée). Nous allons donc réanalyser le fichier en demandant à Fstat de nous donner les F_{IS} par sous-échantillon, puisque nous avons fort judicieusement, il faut bien l'avouer maintenant, d'entrée de jeu distingué les deux sexes.

Sous Fstat vous ouvrez le même fichier "IRTotTestPanmix.dat" et vous décochez toutes les cases et cochez celle qui indique "Fis" dans le cadre "Per locus and sample statistics" comme indiqué dans la figure 18. Si vous souhaitez repérer encore une fois les noms des sous-échantillons, n'oubliez pas de signaler à nouveau l'existence du fichier "IRTotTestPanmix.lab" dans le menu "Options".

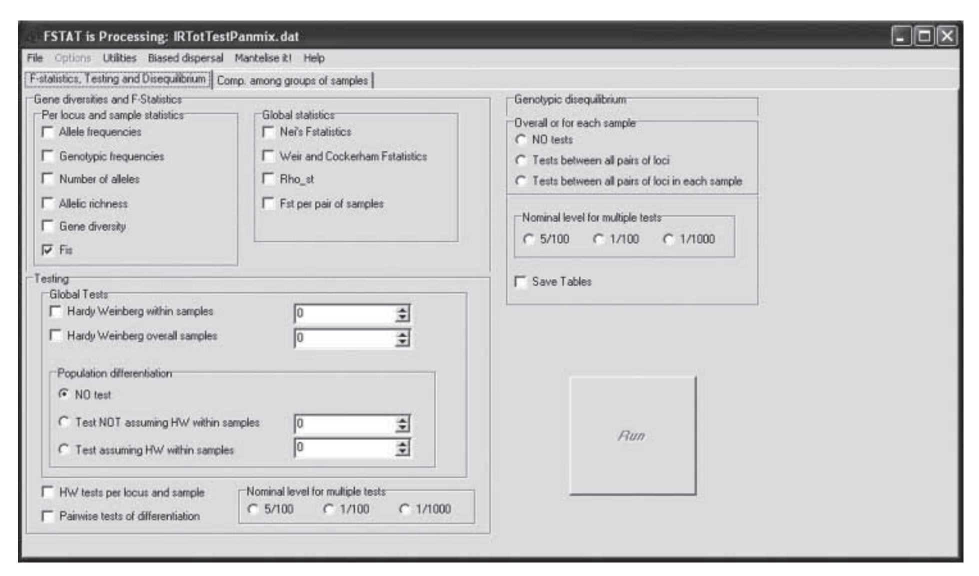


Figure 18
Capture d'écran de Fstat lors de la deuxième analyse.

Quand vous lancerez "Run", Fstat ouvrira une boîte de dialogue avec laquelle vous pouvez décider d'écrire les résultats de cette analyse dans un nouveau fichier. Dans le cas contraire, et c'est le choix que j'ai fait, le programme écrira les résultats dans "IRTotTestPanmix.out" à la suite des analyses précédentes (fin du fichier). Qu'y découvrons-nous ? Tout d'abord que Fstat tronque les labels plus longs que six

caractères. Ce n'est pas grave, car nous avons toujours le même ordre Femelles Mâles pour chaque échantillon. Et puis il suffit (sous Excel c'est facile) de faire un copier-collage spécial/transposition à partir du fichier "IRTotTestPanmix.lab". Ensuite, comme représenté dans la figure 19, construite à partir du fichier de sortie, nous pouvons voir, qu'à part pour le locus IR08, aucune tendance claire n'apparaît. Tous ces loci présentent des déficits importants et relativement variables, mais sans lien réel avec le sexe des tiques. Ce seraient plutôt les mâles qui auraient une tendance à présenter des déficits d'hétérozygotes plus importants (nous verrons plus loin une explication possible). Pour le locus IR08 par contre, avec un $F_{IS} = 1$ pour les mâles, il apparaît clairement que ce locus est situé sur le chromosome X et qu'il est donc haploïde chez les individus mâles.

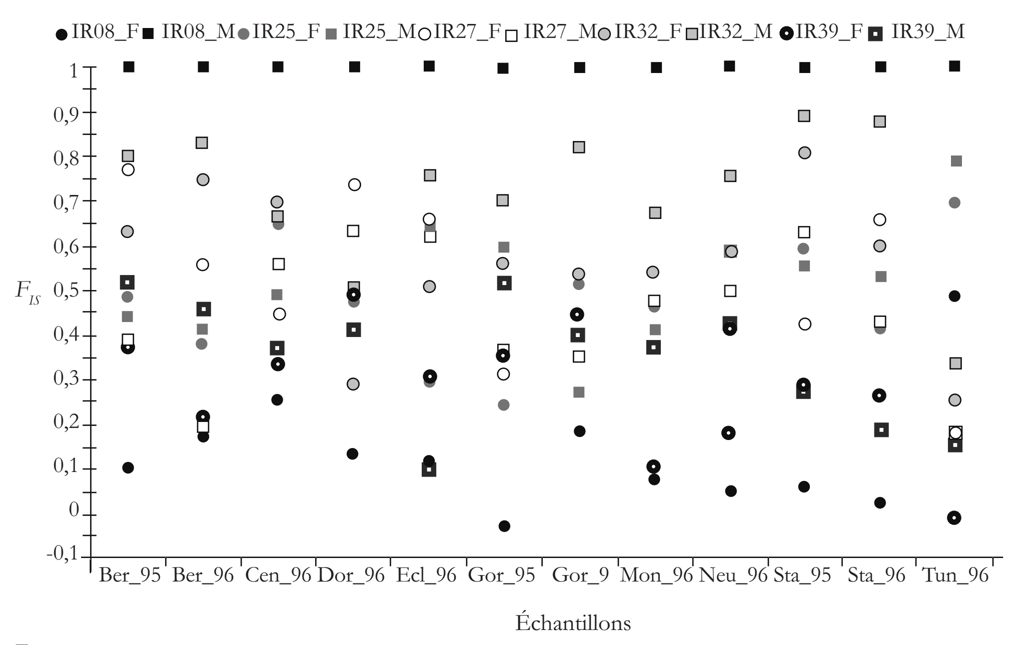


Figure 19
Estimations des F_{IS} par locus et par sous-échantillon. Les abréviations des échantillons sont identiques à celles de la figure 15. Les échantillons de femelles sont représentés par des ronds et ceux des mâles par des carrés.

En fait, pour être précis, le locus IR08 avait été trouvé hétérozygote pour quatre individus mâles sur l'ensemble du jeu de données. Même si cela pouvait refléter des duplications toujours possibles (comme évoqué p. 118), nous avons choisi d'éliminer ces individus, car ils pouvaient correspondre à des erreurs de manipulations.

Quoi qu'il en soit, il va donc falloir recoder les données à ce locus. Pour l'analyse des F_{IS}, les mâles devront en effet être codés en données manquantes (000000) au

locus IR08. Nous allons donc créer un nouveau fichier "IRTotTestPanmixMalManqIR08.dat" à partir du précédent et refaire l'analyse globale du F_{IS}. Celle des déséquilibres de liaison, qui est un test génotypique, n'a aucune raison d'avoir été affectée par ce phénomène. Dans Fstat, nous cocherons donc les mêmes cases qu'en figure 16, à l'exception de celles concernant les déséquilibres de liaison.

Dans le fichier de sortie "IRTotTestPanmixMalManqIR08.out", nous constatons l'image suivante (voir aussi la figure 20) : rien ne change sauf pour le locus R08 qui montre les plus basses valeurs de F_{IS}, mais qui restent très significativement (toutes les *P*-values sont inférieures ou égales au minimum possible 0,0001) au-dessus de la valeur nulle attendue sous panmixie. Notez au passage que je ne me sers des intervalles de confiance que pour illustration. Le F_{IS} global reste donc très élevé (0,39), inexplicablement variable entre loci et fort variable d'un site à l'autre. Ceci suggère un rôle possible pour des allèles nuls ou de dominance d'allèles courts. L'étape suivante sera donc de mettre en évidence l'existence de tels allèles et/ou de phénomène de dominance.

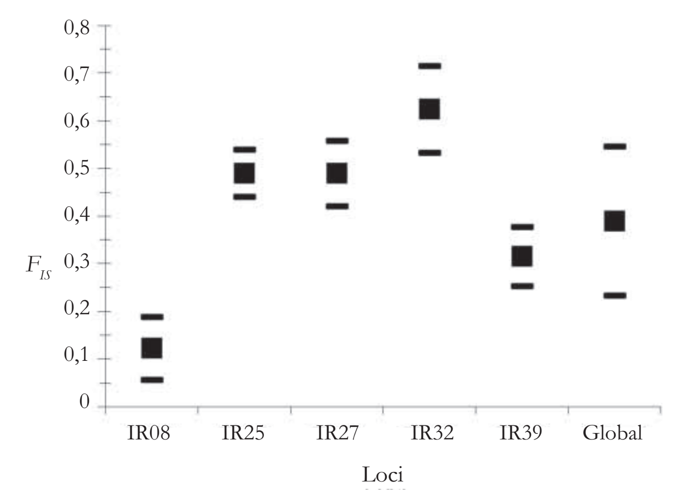


Figure 20
Valeurs moyennes du F_{IS} par locus et intervalle de confiance obtenus pour les microsatellites d'*Ixodes ricinus*, avec les mâles codés comme données manquantes pour le locus IR08. Pour chaque locus, les intervalles de confiance sont calculés à l'aide de l'erreur standard donnée par le jackknife sur les populations et la valeur du *t* pour 11 ddl (12 – 1) (la moitié des échantillons) et α = 0,05 (soit 2,201, voir le tableau 3) en suivant l'équation (45). Pour la valeur globale, l'intervalle de confiance est issu du bootstrap sur les loci.

RECHERCHE D'ALLÈLES NULS ET DE DOMINANCE D'ALLÈLES COURTS

Nous allons pour ce faire utiliser deux nouveaux logiciels. Micro-Checker va nous permettre d'estimer la fréquence des allèles nuls susceptibles d'expliquer, dans chaque sous-échantillon et pour chaque locus, les déficits en hétérozygotes observés. Micro-Checker permet également d'estimer si les données sont compatibles avec un bégaiement de la polymérase (*stuttering*) et/ou une dominance des allèles les plus courts. Pour la dominance des allèles courts, nous utiliserons également une méthode plus puissante que celle implémentée par Micro-Checker. Nous allons procéder à une régression généralisée pour la mise en œuvre de laquelle nous utiliserons le logiciel R (voir la référence complète dans la bibliographie).

Convertir le fichier pour Micro-Checker et ouverture du logiciel

Pour commencer avec Micro-Checker, nous avons besoin de transformer nos données au format Genepop qui est compatible avec ce logiciel. Ensuite, nous allons devoir créer un fichier spécial pour les données du locus IR08, lié au sexe, sans les mâles car sinon Micro-Checker risque de goûter moyennement la saveur de cette plaisanterie. Créons donc un fichier "IR08AllFem.txt" avec les données femelles pour le seul locus IR08 et un fichier "IRAutosomAll.txt" pour le reste des données. Attention, le fichier doit suivre des règles strictes sinon Micro-Checker refusera d'analyser les données. Référez-vous au fichier exemple fourni avec le logiciel et respectez les espaces et tabulations de la façon la plus scrupuleuse (ou utilisez Create).

Lancez Micro-Checker et ouvrez "IRAutosomAll.txt" avec le menu "File". Si tout se passe bien, vous observez l'ouverture de votre fichier avec vos données et différents menus et boutons en bas de l'écran.

Analyses des loci autosomiques du premier sous-échantillon par Micro-Checker

Il y a un encadré en bas à gauche où il faut choisir le motif de chaque locus microsatellite. Il affiche par défaut le premier des loci (ici IR25) et un blanc pour le motif. Choisissez le motif "Mononucleotide" comme sur la figure 21.

Nous avons déjà remarqué que nos loci microsatellites étaient peu orthodoxes. L'option mononucléotidique correspond en fait à l'option qui permet de faire face à toutes les situations. Cliquez ensuite sur le bouton "All" pour signaler que cette option est valable pour tous les loci. Cliquez ensuite sur le bouton "Analyse" (un peu

	17	Ber_95_F	000	000	123	123	233
	18	Ber_95_F	130	130	119	119	233
	19	Ber_95_F	145	145	119	121	243
	20	Ber_95_F	147	147	119	119	243
	21	Ber_95_F	000	000	119	121	233
	22	Ber_95_F	150	151	119	119	233

Enter the Repeat Motif for each Locus All
IR25 Mononucleotide
Maximum Expected Allele Size 350
Unusual Observations Check
Confidence Interval 95% Analyse 1000 x
Next Population
Loaded Pops.
Exit
Ready

Figure 21
Capture d'écran de Micro-Checker.

plus à droite). Apparaît alors une fenêtre d'avertissement comme celle présentée en figure 22. Comme il y a des données manquantes, Micro-Checker vous demande s'il faut ou non en tenir compte. Autrement dit, les données manquantes correspondent-elles à des homozygotes nuls (blancs) et faut-il les utiliser pour le calcul des fréquences des allèles nuls par la seconde méthode de Brookfield (1996) ? La réponse étant positive, cliquez donc directement sur “Proceed” sans vous poser plus de questions.

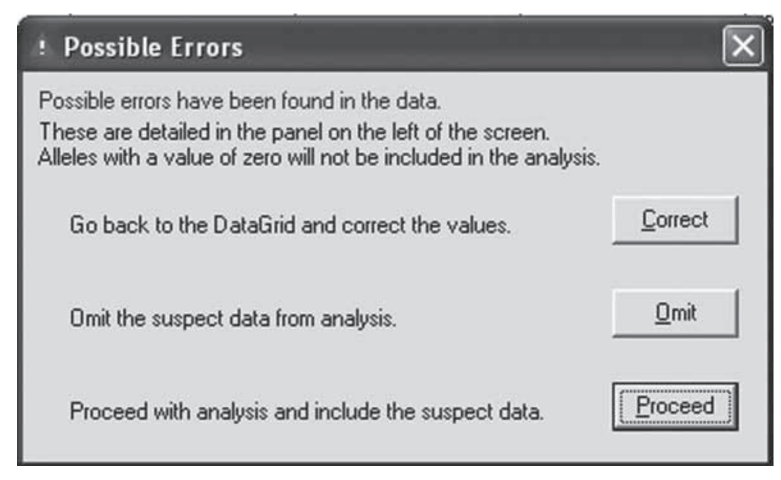


Figure 22
Cadre d'invite de commande de MicroChecker pour définir la nature des données manquantes et s'il faut en tenir compte dans le calcul des fréquences des allèles nuls.

Micro-Checker effectue plusieurs calculs et vous présente des résultats concernant le premier locus. Allez dans le menu “Tools” à “Nulls across loci” comme dans la figure 23 pour obtenir le tableau des fréquences de nuls dans le premier sous-échantillon, estimées selon différentes méthodes. Sélectionnez ce tableau avec la souris, copiez-le et sauvez-le dans un fichier (Excel, par exemple). Ensuite, regardez dans l’encadré en bas à droite (fig. 23) si le locus correspondant montre un problème de stuttering ou une dominance d’allèle court (« large allele dropout »). Si oui, notez-le dans le tableau que vous venez de créer pour sauvegarder les résultats de cette analyse puis, par le menu “Window’” (fig. 23) sélectionnez le locus suivant, etc. Vous constaterez qu’aucun locus ne présente de « stuttering » ni de dominance d’allèle court dans ce premier sous-échantillon.

Analyses des autres sous-échantillons, des autres loci autosomiques et du locus IR08

Au centre et en bas, cliquez sur le bouton “Next Population” (voir fig. 23) pour analyser le sous-échantillon suivant en reprenant les mêmes étapes décrites en p. 125-127, jusqu’au dernier sous-échantillon. N’oubliez pas de copier le tableau des fréquences d’allèles nuls à chaque fois (dans le menu “Tools” à “Nulls across loci”, fig. 23). Ensuite, vous ferez la même chose pour le locus lié au sexe, IR08, en ouvrant le fichier correspondant “IR08AllFem.txt”.

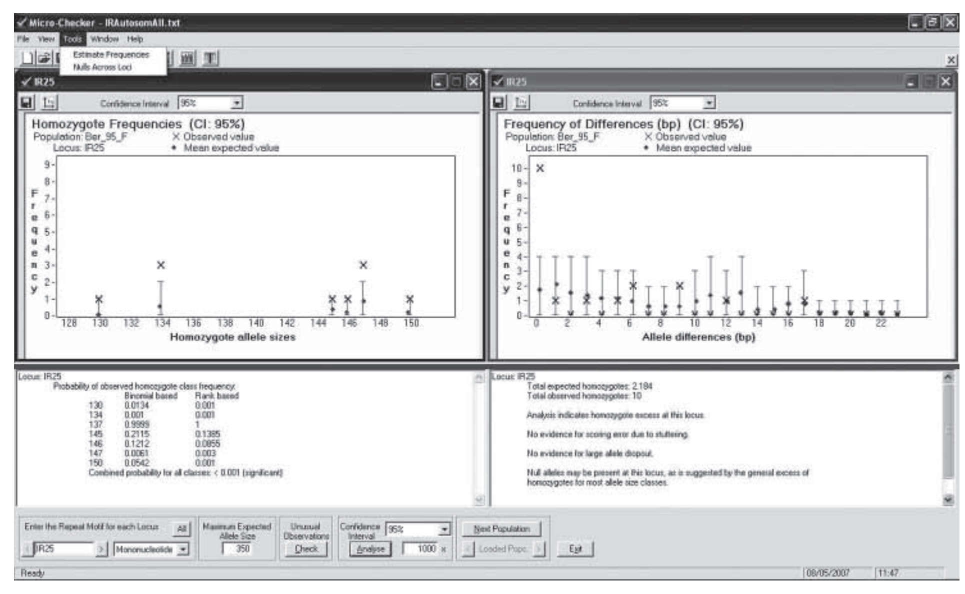


Figure 23
Sortie de MicroChecker vous indiquant, pour le locus et le sous-échantillon mentionné, la présence ou non de problèmes.

Bilan des analyses avec Micro-Checker

Nous avons constitué un fichier de résultats avec les fréquences d'allèles nuls probables, l'existence ou non de stuttering et de dominance d'allèles courts. Nous ne gardons que la méthode 2 de Brookfield (1996) qui tient compte des données manquantes (blancs) comme des homozygotes nul/nul. Dans ce fichier, nous allons également insérer le nombre d'individus génotypés pour chaque locus (copiés à partir des fichiers de sortie Fstat), la fréquence attendue sous panmixie (fréquence précédente au carré) des allèles nuls pour chaque locus dans chaque sous-échantillon et sur l'ensemble des sous-échantillons, le nombre de blancs observés (compter les 000000 dans chaque sous-échantillon et sur l'ensemble), l'effectif corrigé (individus génotypés + blancs) et enfin le nombre de blancs attendus sous la double hypothèse qu'il y a panmixie et que les allèles nuls expliquent les F_{IS} en totalité. Le tableau 8 donne un aperçu du fichier final pour le locus IR08.

Tableau 8
Synthèse des résultats de Micro-Checker pour le locus IR08 chez les femelles *Ixodes ricinus*. La fréquence attendue des blancs $p_{B2}{}^2$ est obtenue en mettant au carré la fréquence estimée des allèles nuls selon la méthode 2 de Brookfield (1996) et le nombre de blancs attendus correspondant à cette valeur multipliée par N'. N' correspond, quant à lui, à la somme de N (individus génotypés) et des blancs observés. Pour la dernière ligne, la valeur de $p_{B2}{}^2$ est obtenue en divisant le nombre total de blancs attendus par le N' total.

Sous-échantillon	Nul	Stuttering	Brookfield 2	$p_{B2}{}^2$	N	N'	Blancs observés	Blancs attendus
Ber_96_F	oui	non	0,1201	0,0144	45	46	1	0,66
Cen_96_F	oui	non	0,1736	0,0301	29	30	1	0,90
Dor_96_F	oui	non	0,0594	0,0035	47	47	0	0,17
Gor_96_F	oui	oui	0,0826	0,0068	43	43	0	0,29
Tun_96_F	oui	non	0,3594	0,1292	18	20	2	2,58
Tous				0,0253	182	186	4	4,61

Pour vérifier que ces résultats expliquent correctement les F_{IS} observés, on peut comparer la proportion de blancs observés avec celle attendue sous l'hypothèse que les allèles nuls expliquent la totalité de ces F_{IS}. Un test binomial unilatéral avec comme fréquence attendue $p_{B2}{}^2$, un nombre de réussite égal aux blancs observés pour un nombre d'essais de N', semble ici approprié. On préfère ici un test unilatéral, car ce qui nous intéresse est de savoir si on a oui ou non moins de blancs qu'attendus. On peut facilement effectuer ce test sous R.

Il nous faut donc lancer R et dans la fenêtre de commande taper l'instruction :

```
binom.test(Blancs observés, N□, p = pB2², alternative = □less□)
```

Pour des raisons de recherche de puissance et pour limiter le nombre de tests dont la multiplication est toujours problématique (voir p. 82 en première partie), on ne fera les tests qu'avec les valeurs totales pour chaque locus. Pour le locus IR08, cela correspond aux valeurs de la dernière ligne du tableau 8. Pour ce locus, la commande devient donc :

```
binom.test(4, 186, 0.0253, alternative="less")
```

Faites bien attention de respecter strictement le format (en particulier, les majuscules et minuscules sont reconnues comme des caractères différents sous R). Ici "less" signifie que le test est unilatéral dans le sens des plus petites valeurs (H1 : il y a moins de blancs observés qu'attendus) (l'instruction devient "two.sided" pour un bilatéral et "greater" pour l'autre test unilatéral). Une fois que vous avez tapé cette instruction dans R, tapez sur la touche "Entrée" et le test se fait. La *P*-value du test est, pour IR08, non significative (*P*-value = 0,4919). Les allèles nuls sont donc bien suffisants pour expliquer les déficits en hétérozygotes observés à ce locus chez les femelles, d'autant plus qu'il semble aussi exister des phénomènes de stuttering à ce locus. Pour les autres loci, on procède de la même façon. On trouve ainsi que pour les loci IR 25, IR27 et IR32, la fréquence des blancs observés est significativement inférieure à celle des blancs attendus si les allèles nuls devaient expliquer les déficits en hétérozygotes. C'est un problème car, par un phénomène de cercle vicieux, moins les allèles nuls expliquent un déficit en hétérozygotes, moins le nombre de blancs observés correspond aux attendus. Pourquoi cela ? Simplement parce que si on attend naturellement plus d'homozygotes en général, alors on devrait observer encore plus d'homozygotes nuls (blancs), en particulier (ce raisonnement ne marche cependant pas très bien s'il s'agit d'un effet Wahlund). Par ailleurs, la variance entre loci ainsi que le fait que les nuls expliquent très bien les déficits observés pour IR08 (voir plus haut), mais aussi pour IR39 (*P*-value = 0,312) pourraient nous inciter à exclure des causes biologiques du type régime de reproduction ou effet wahlund (voir plus loin). Notons que des phénomènes de stuttering ont été détectés pour IR25, mais seulement dans deux sous-échantillons. Pour IR32 et IR27, Micro-Checker n'a pas détecté ce phénomène pas plus qu'il n'a détecté de dominance d'allèles courts. Cependant, Micro-Checker ne travaille que dans chaque sous-échantillon de façon isolée, ce qui peut représenter une forte perte de puissance. Dans le paragraphe qui suit, nous allons utiliser une autre technique pour détecter d'éventuelles dominances d'allèles courts.

Détection de dominance d'allèles courts par la méthode de régression multiple

Pour ce faire, nous aurons besoin de connaître, pour chaque locus et dans chaque sous-échantillon, la valeur du F_{IS} pour chaque allèle. On peut demander à Genetix de le faire en choisissant à chaque traitement le locus et le sous-échantillon à

analyser, en n'oubliant pas de zapper les mâles au locus IR08. On peut aussi créer autant de fichiers Fstat qu'il y a de sous-échantillons à analyser, ensuite, et parce que malheureusement Fstat ne permet pas d'analyser qu'un seul sous-échantillon, il faut créer dans chaque fichier une deuxième population fictive, de taille identique à celle à analyser et fixée à tous les loci (par exemple, tous homozygotes 170170, 150150, 123123, 235235, 129129 pour les cinq loci respectivement). Il s'agit ensuite de récupérer dans chaque sous-population les F_{IS} de chaque allèle pour chacun des cinq loci et de créer cinq fichiers de données (un par locus) contenant pour chaque allèle son F_{IS}, sa taille (on s'en doute), le sous-échantillon, sa fréquence allélique p dans ce sous-échantillon, le produit $p(1\text{-}p)$, le nombre d'individus génotypés dans ce sous-échantillon N et enfin le produit $p(1\text{-}p)N$. Le tableau 9 donne une idée de la forme de ce fichier pour le locus IR08 que j'ai appelé "IRTotL08MalManqFisAllSizeL08.txt". Pour fabriquer ce fichier, une feuille de calcul Excel est idéale, ensuite il suffit d'enregistrer le fichier en format texte seul.

On peut aussi utiliser Genetix qui permet l'analyse d'un seul sous-échantillon, mais dont les sorties sont moins commodes à importer dans Excel (à vous de voir).

La colonne Npq, qui donne en fait le résultat du produit $Np(1\text{-}p)$, nous servira à pondérer notre régression par la taille des échantillons, mais en donnant aussi plus de poids aux allèles de fréquences proches de 0,5 (les plus polymorphes). On fait les mêmes fichiers avec les quatre autres loci. Nous allons maintenant analyser ces données avec le logiciel R.

Tableau 9
Aperçu du fichier de données pour le locus IR08 en vue de l'analyse de régression du F_{IS} en fonction de la taille des allèles et du sous-échantillon.

F_{IS}	Allele	Sample	Year	Sex	p	N	pq	Npq
- 0,02439	165	Bern	95	F	0,0455	22	0,04342975	0,9554545
0	166	Bern	95	F	0,0227	22	0,02218471	0,48806362
- 0,05	168	Bern	95	F	0,0682	22	0,06354876	1,39807272
0,65574	169	Bern	95	F	0,0682	22	0,06354876	1,39807272
- 0,02439	170	Bern	95	F	0,0455	22	0,04342975	0,9554545

Ouvrez R et dans le menu "Fichier" cliquez dans "Changer le répertoire courant...", et allez dans le répertoire où vous avez stocké vos fichiers de données. Dans la console de travail de R, tapez la suite de commandes, chacune suivie d'un retour chariot (touche "Entrée") :

```
> data<-read.table("IRTotL08MalManqFisAllSizeL08.txt", header=TRUE)
```

qui signifie que le tableau de données “data” est contenu dans le fichier nommé et que la première ligne contient le nom des colonnes. N’oubliez pas que les données manquantes se notent “NA” en majuscules et non “000000”.

```
> attach(data)
```

qui signifie que ce tableau doit être chargé en mémoire[9].

```
> loc8<-glm(data, formula = Fis ~ poly(Allele, 2) + Sample + Year, family
= gaussian, weights = Npq)
```

où loc8 est le nom d’un modèle linéaire généralisé utilisant le tableau “data” et dont la régression tente d’expliquer la valeur du F_{IS} en fonction de la taille des allèles selon un polynôme d’ordre 2 ou quadratique (qui s’est avérée plus proche de ce qui se passe dans le cas qui nous intéresse), du sous-échantillon d’origine et de l’année. Le sexe n’a ici aucune importance puisqu’il n’y a que des femelles. Nous ne testons l’effet d’aucune interaction entre variable, car en fait je ne vois aucune raison pour qu’il en existe. Pensez à respecter les majuscules s’il y en a, car R les reconnaît comme telles. Tapez enfin :

```
> anova(loc8, test="F")
```

qui renvoie à une analyse de variance utilisant la statistique F (se référer à un livre de statistique pour approfondir ces notions) et donne le résultat suivant :

```
Analysis of Deviance Table
Model: gaussian, link: identity
Response: Fis
Terms added sequentially (□rst to last)
                Df  Deviance   Resid. Df  Resid. Dev F       Pr(>F)
NULL                           198        21.6160
polyAllele, 2) 2   0.4021     196        21.2139    2.1174  0.1232242
Sample          8   3.1604     188        18.0536    4.1609  0.0001339 ***
Year            1   0.2995     187        17.7540    3.1550  0.0773192
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Ici, on voit que seul le sous-échantillon influence la valeur du F_{IS} (allèles nuls, stuttering variable dans l’espace ?) qui n’explique que 14,63 % de la dispersion (100×3,1604/21,616), tout en étant très significatif.

On utilise un test F, car on a supposé que la distribution des F_{IS} suit plus ou moins une courbe de Gauss (données continues en cloche symétrique), ce qui est sûrement inexact mais ne risque guère de modifier le résultat dans un sens dramatique.

Pour les loci suivants, nous aurons besoin de distinguer le sexe des tiques.

[9] Entre temps, j’ai découvert l’existence du “Package” R-Commander ou Rcmdr qui, en quelques clics de souris, permet d’effectuer ces commandes automatiquement.

Avec le locus IR25, l'analyse du fichier "IRTotSexSepFisAllSizeL25.txt" est la suivante :

```
> data<-read.table("IRTotFisAllSizeL25.txt",header=TRUE)
> attach(data)
> loc25<-glm(data, formula = Fis ~ poly(Allele, 2) + Site + Year + Sex,
family = gaussian, weights = Npq)
> anova(loc25, test="F")
```

Ce qui aboutit au tableau de résultat :

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			326	60.844		
poly(Allele, 2)	2	0.123	324	60.721	0.3420	0.71062
Sample	8	3.729	316	56.992	2.591	0.00938 **
Year	1	0.160	315	56.832	0.888	0.34675
Sex	1	0.349	314	56.483	1.942	0.16438

On aboutit à une conclusion similaire à la précédente, puisque ni le sexe ou l'année ni la taille des allèles ne comptent avec seulement environ 6,13 % de la déviance expliquée par le site qui est moins spectaculairement significatif que précédemment.

Pour le locus IR27, le tableau obtenu est différent :

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			133	25.9549		
poly(Allele, 2)	2	4.1186	131	21.8363	15.1968	1.294e-06 ***
Sample	8	5.1810	123	16.6553	4.7793	4.022e-05 ***
Year	1	0.0621	122	16.5932	0.4584	0.4997
Sex	1	0.1967	121	16.3964	1.4519	0.2306

En effet, comme nous pouvons le déduire du tableau ci-dessus, le site (Sample) explique 19,96 % de la dispersion des points (5.181/25.9549) et la taille des allèles (poly(Allele, 2)) en explique 15,86 % (4.1186/25.9549) et sont tous les deux très significatifs (souligné par les trois étoiles). Ils expliquent ainsi 35,83 % de la variance. Cette valeur est conséquente eu égard à l'importante variance résiduelle attendue en général pour un estimateur de statistique *F*. Comme le montre la courbe décrite dans la figure 24, la relation entre F_{IS} et taille des allèles est négative (si on exclut les trois premiers points, ce qui ne changerait rien eu égard aux intervalles de confiance), ce qui peut donc être interprété par une dominance des allèles les plus courts.

Pour le locus IR32, on observe le résultat suivant :

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			191	38.762		
poly(Allele, 2)	2	0.340	189	38.422	1.1742	0.3114224
Sample	8	10.155	181	28.267	8.7720	4.318e-10 ***
Year	1	0.089	180	28.178	0.6156	0.4337072
Sex	1	2.275	179	25.903	15.7179	0.0001062 ***

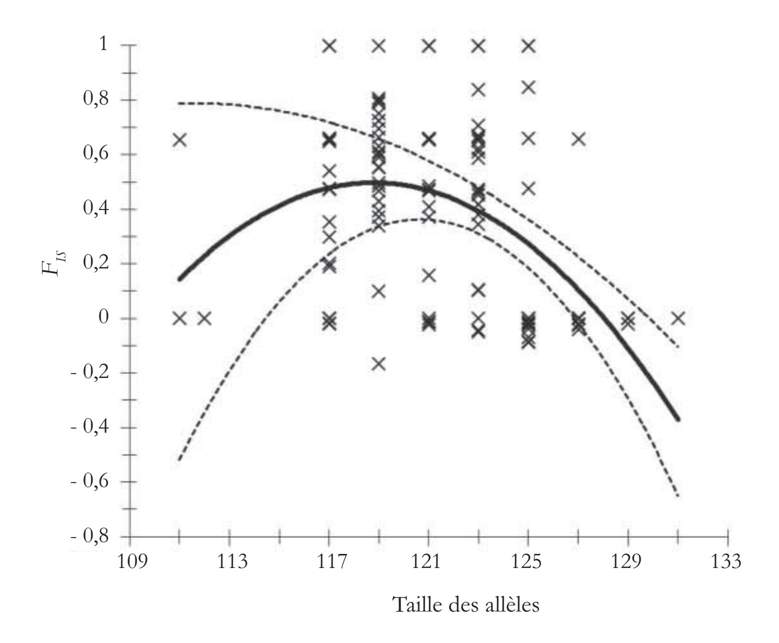


Figure 24
Relation entre taille des allèles et F_{IS} pour le locus IR27 et sur l'ensemble des échantillons.
Les intervalles de confiance à 95 % ont été obtenus avec $F_{IS} \pm t_{0,05;N1} \times \sqrt{\frac{Variance(F_{IS})}{N}}$.
Pour ce faire, les singletons (tailles d'allèles présents une seule fois comme 112 et 131) ont été réunis à la classe la plus proche.

On voit qu'en plus du site, le sexe des tiques a un effet significatif, ce qui signifie que nous avons eu raison d'en tenir compte et nous verrons ensuite pourquoi.

Pour le locus IR39, le tableau obtenu est le suivant :

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			368	59.156		
poly(Allele, 2)	2	0.932	366	58.223	3.2447	0.04013 *
Sample	8	6.139	358	52.084	5.3426	2.372e-06 ***
Year	1	.419	357	51.665	2.9159	0.08858
Sex	1	0.529	356	51.136	3.6804	0.05585

Le site joue une fois encore de façon significative, mais aussi la taille des allèles, même si cette dernière n'explique même pas 2 % de la déviance et est peu significative. Par ailleurs, la figure 25 montre que la relation (augmentation globale du F_{IS} avec la taille des allèles) n'est pas compatible avec une dominance des allèles courts. On peut donc attribuer ce résultat au hasard et au nombre de tests effectués qui augmente la probabilité d'obtenir quelque chose de significatif par

hasard (revoir la première partie de ce manuel, p. 82-87). Rappelons que pour ce locus, les allèles nuls s'étaient avérés suffisants pour expliquer les déficits en hétérozygotes observés. Il est plus raisonnable ici de considérer ce résultat comme fortuit.

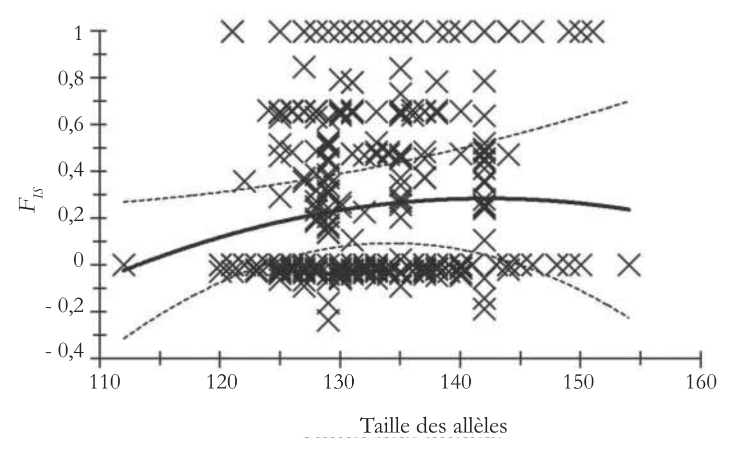


Figure 25
Relation entre F_{IS} et la taille des allèles au locus IR39 sur l'ensemble des échantillons.
Les intervalles de confiance à 95 % ont été obtenus comme précédemment.
Les sous-échantillons de moins de quatre individus ont été associés au plus proche.

Bilan de l'analyse des déficits locaux en hétérozygotes

Pour les loci IR08 et IR39, les allèles nuls semblent pouvoir expliquer les forts et variables F_{IS} observés. Pour IR27, les allèles nuls et la dominance des allèles courts offrent conjointement une explication satisfaisante. Seul le locus IR32 offre des déficits énormes et non expliqués par les allèles nuls, le « stuttering » ou la dominance des allèles courts. Cependant, sachant que le « stuttering » n'a pu être testé que sous-échantillon par sous-échantillon (manque de puissance), que la plupart des allèles se suivent à un pas sur ce locus et compte tenu de ce que nous trouvons aux autres loci, il est possible qu'ici aussi les déficits observés proviennent d'un problème technique.

Je peux ajouter ici qu'un module (package) de R, appelé "R-Commander", dont je n'ai appris l'existence qu'après la rédaction de ce chapitre, permet d'accéder aux analyses effectuées dans ce paragraphe à l'aide de menus déroulants plus conviviaux que le mode commande strict.

RECHERCHE D'UNE STRUCTURE CACHÉE (EFFET WAHLUND)

Introduction

Nous allons dans un premier temps continuer de considérer les femelles et les mâles séparément. On sait en effet qu'il y a une structure génétique spécifique pour chaque sexe dans ce jeu de données. Même si nous analyserons ceci plus tard, il n'est pas inutile de poursuivre la recherche d'explications des déficits en hétérozygotes avant d'aborder cet aspect. Nous allons donc analyser tous les sous-échantillons (mâles et femelles séparées) pour obtenir l'information sur le plus grand nombre de réplicas possibles. Ensuite, nous nous concentrerons sur 1996 en réunissant les mâles et les femelles pour faire des tests.

Le but du jeu sera ici d'utiliser l'information multilocus de chaque individu, dans chaque sous-échantillon afin de vérifier à l'aide du logiciel BAPS (voir p. 98-101 en première partie et le tableau 1 en annexe), si certains individus peuvent être regroupés sur la base de leur ressemblance génétique. BAPS va ensuite explorer de façon itérative et répétée, en suivant plusieurs chaînes de Markhov (ou une chaîne stochastique d'optimisation suivant les versions) afin de trouver la meilleure partition (celle qui regroupe le mieux les individus) dans le sous-échantillon analysé. La partition définit un nombre donné de clusters (sous-unités) composés chacun d'un certain nombre d'individus du sous-échantillon. La qualité d'une partition se définit par un savant calcul dans le détail duquel je serai bien incapable de rentrer, mais qui dépend de la distance génétique entre les groupes définis, par rapport aux autres partitions explorées durant le processus. Il est aussi expliqué, dans les articles décrivant le logiciel, qu'une hypothèse du modèle utilisé dans l'algorithme est que les « clusters » qui composent la partition sont en équilibre de Hardy-Weinberg. Je ne suis pas certain de bien comprendre ce qui est entendu par là dans la mesure où mon expérience m'a montré que la plupart des partitions obtenues ne sont pas conformes à cet équilibre, voire même en sont très éloignées. J'ai également pu observer cela avec STRUCTURE qui fait la même hypothèse. Comme discuté dans la première partie de ce manuel, beaucoup reste à explorer concernant le fonctionnement de ces méthodes dans différentes situations. Il faudra donc vérifier si la partition obtenue (car le logiciel en donne toujours une) correspond à quelque chose de viable et pas seulement une vue de l'esprit.

Si la partition a réellement mis en évidence des groupes cryptiques au sein des sous-échantillons susceptibles d'expliquer en partie (effet Wahlund) nos fameux déficits en hétérozygotes, il faudra ensuite trouver et explorer les hypothèses susceptibles d'expliquer le plus raisonnablement possible (mais en aveugle) ces résultats (espèces

ou races d'hôtes cryptiques, sous-structures familiales, isolement par la distance entre individus sur de courtes distances).

Il existe d'autres logiciels qui en principe font la même chose. L'avantage de BAPS réside dans sa convivialité, dans le fait qu'il accepte des fichiers de type Genepop (un peu modifiés) et qu'il m'a toujours donné de bons résultats. Le logiciel STRUCTURE est par exemple beaucoup moins commode à utiliser (et c'est un euphémisme) et, sur un même jeu de données (glossines), n'a pas offert de partitions aussi satisfaisantes que BAPS (Ravel *et al.*, 2007). Des études comparatives de différents logiciels de clustering sont en cours, mais la longueur et la quantité des analyses font que des résultats concrets ne seront sans doute pas disponibles avant la sortie du présent ouvrage. Vous verrez aussi l'application d'un autre logiciel de même nature, Flock, plus loin dans cette partie.

Construction des fichiers BAPS

Il faut construire un fichier pour chaque sous-échantillon. Le type est semblable à un fichier Genepop, mais avec des tabulations comme dans la figure 26 (symbolisées par des →) qui donne un exemple pour le fichier des mâles de Staadswald. On note que les mâles sont codés homozygotes pour IR08 afin que l'information multilocus soit préservée pour cinq loci. Par commodité, j'ai appelé ce fichier "IRTotBrut1Stad95M.gen", mais vous faites comme bon vous semble.

Ensuite, il est commode de créer un fichier texte contenant le chiffre 30 répété un grand nombre de fois (ici 50 fois), avec un espace entre chaque répétition et sur une seule ligne. Le logiciel BAPS vous demandera en effet de taper un nombre maximal probable pour les clusters. Ici, 30 m'est apparu comme largement raisonnable compte tenu des tailles de sous-échantillons. C'est à partir de ce chiffre que BAPS démarre et recherche une partition la plus probable en se limitant à ce nombre maximum de clusters. Le logiciel reprend ensuite le processus autant de fois que l'on a rentré ce chiffre (ici 50) et ne gardera que la meilleure de toutes les partitions explorées. Avoir tapé 50 fois ce chiffre dans un fichier permet de copier et coller cette séquence directement sans avoir à la retaper pour toutes les analyses. J'ai appelé ce fichier "50fois30.txt" (quelle imagination !).

Analyse des fichiers par BAPS

Vous avez bien entendu installé BAPS sur votre machine et créé tous les fichiers nécessaires (il y en 24 normalement). Il faut maintenant lancer BAPS en cliquant sur BAPS4_RUNME.EXE. Le logiciel ouvre deux fenêtres, une fenêtre Dos dont il n'est pas vraiment nécessaire de se préoccuper maintenant et une fenêtre d'interface type Windows avec des menus que nous allons utiliser. Il est important de commencer par créer un fichier résultat. Pour ce faire, cliquez sur "File", "Output File" et "Set"

```
Sta95M¶
IR08¶
IR25¶
IR27¶
IR32¶
IR39¶
Pop¶
Staa95M → , → 000000 → 000000 → 119119 → 000000 → 127134¶
Staa95M → , → 174174 → 136147 → 119119 → 246246 → 132132¶
Staa95M → , → 170170 → 137144 → 119119 → 241250 → 000000¶
Staa95M → , → 174174 → 000000 → 119121 → 233233 → 135135¶
Staa95M → , → 178178 → 000000 → 119119 → 233233 → 130140¶
Staa95M → , → 169169 → 148151 → 119119 → 233233 → 129142¶
Staa95M → , → 178178 → 134134 → 119123 → 000000 → 125142¶
Staa95M → , → 186186 → 134141 → 119119 → 000000 → 142142¶
Staa95M → , → 175175 → 145145 → 119119 → 233233 → 130130¶
Staa95M → , → 183183 → 144148 → 117117 → 250250 → 128140¶
Staa95M → , → 176176 → 142145 → 119119 → 233233 → 128128¶
Staa95M → , → 177177 → 000000 → 123123 → 246250 → 130139¶
Staa95M → , → 170170 → 148148 → 123123 → 250250 → 140140¶
Staa95M → , → 000000 → 146146 → 119119 → 233233 → 129131¶
Staa95M → , → 178178 → 153153 → 119119 → 235235 → 135135¶
Staa95M → , → 183183 → 142144 → 119119 → 254254 → 135135¶
Staa95M → , → 173173 → 144144 → 119123 → 243243 → 125136¶
Staa95M → , → 170170 → 152152 → 119119 → 250250 → 131134¶
Staa95M → , → 174174 → 151151 → 119119 → 233233 → 125142¶
Staa95M → , → 168168 → 134134 → 119119 → 233233 → 129142¶
Staa95M → , → 174174 → 145147 → 119119 → 233233 → 128142¶
Staa95M → , → 168168 → 145145 → 119119 → 233233 → 128134¶
Staa95M → , → 174174 → 000000 → 117119 → 000000 → 142142¶
Staa95M → , → 171171 → 137137 → 123123 → 236236 → 128135¶
Staa95M → , → 173173 → 134149 → 119123 → 243243 → 130133¶
Staa95M → , → 171171 → 134134 → 123123 → 250250 → 128133¶
Staa95M → , → 179179 → 134134 → 121121 → 243243 → 125142¶
Staa95M → , → 169169 → 000000 → 119119 → 235235 → 130131¶
```

Figure 26
Format de fichier pour BAPS pour les tiques mâles du Staadswald en 1995.
Le locus IR08 est codé homozygote chez les mâles.

et créez un fichier en tapant son nom et en le plaçant dans le répertoire qui vous convient le mieux (là où sont vos données) (voir fig. 27).

Il vaut mieux garder un nom de fichier qui permette de retourner ensuite au fichier de données correspondantes. Ici, le premier fichier analysé sera "IRTotBrut1Ber95F.gen" (femelles de Berne 1995), je choisis donc ici de nommer et créer le fichier résultat "IRTotBrut1Ber95FBAPSRes.txt". Ensuite, il faut cliquer sur le bouton "Clustering of individuals" (fig. 27). Apparaît alors une nouvelle fenêtre de dialogue qui vous propose différents formats de fichiers de données (fig. 28). Choisissez bien entendu le format Genepop en cliquant sur le bouton correspondant. Une fenêtre qui s'ouvre vous permet de naviguer vers le répertoire où se trouve IRTotBrut1Ber95F.gen que vous sélectionnez (soit en tapant son nom complet, soit en tapant *.gen et

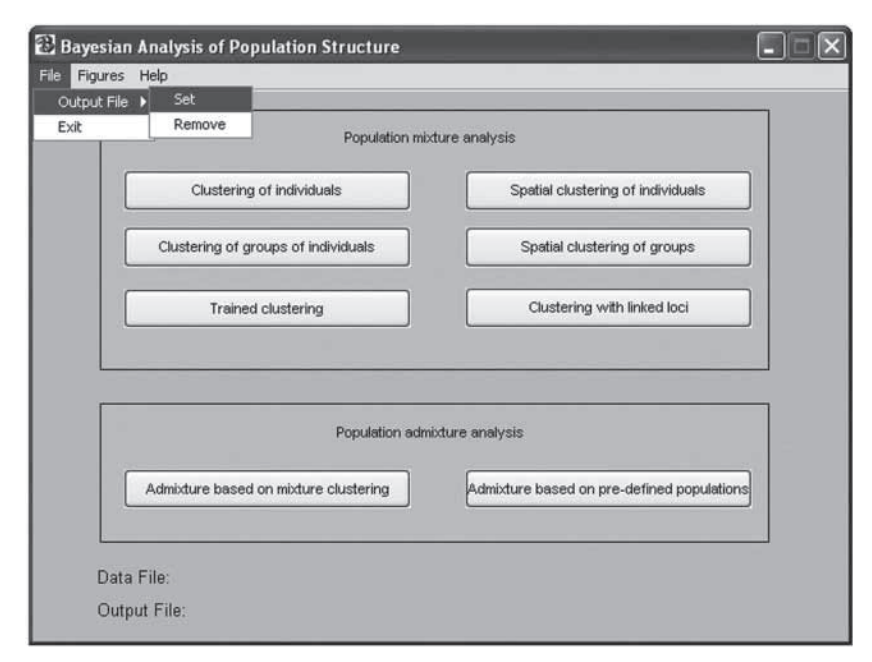


Figure 27
Sélection dans BAPS du fichier de résultats.

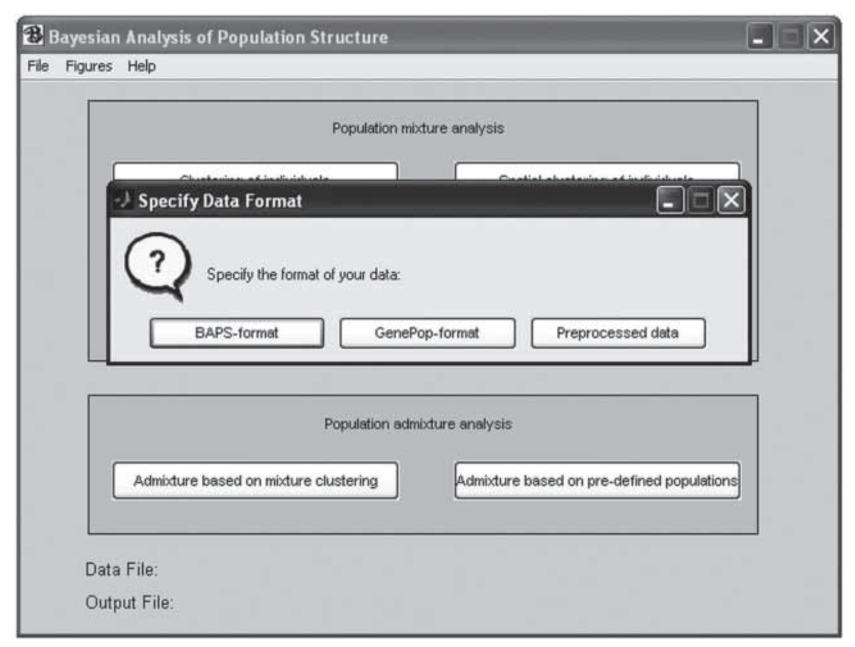


Figure 28
Sélection dans BAPS du format de fichier de données à analyser.

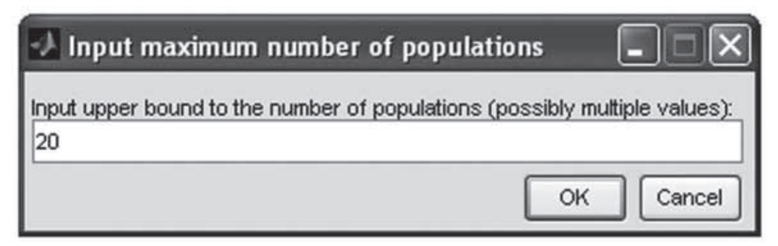


Figure 29
Fenêtre de sélection de la taille maximale des clusters et du nombre de chaînes d'itérations.

retour chariot et en double cliquant sur le fichier). Une question vous est alors posée "Do you wish to save pre-processed data?", cliquez sur "No".

C'est alors qu'apparaît une petite fenêtre permettant de sélectionner le nombre maximum de clusters, ainsi que le nombre de chaînes d'itérations à effectuer (fig. 29), comme expliqué en p. 136. Supprimez le chiffre par défaut (20) et remplacez-le par la chaîne de 30 que vous copiez à partir de "50fois30.txt", collez cette chaîne dans la case idoine et cliquez sur "OK".

Les calculs démarrent et se poursuivent jusqu'à la fin où la meilleure partition est sauvée dans "IRTotBrut1Ber95FBAPSRes.txt". Apparaissent un graphique censé représenter la partition (clusters de différentes couleurs), dont on ne va pas se servir, ainsi qu'un dialogue final vous demandant si vous souhaitez sauver ces données en vue d'une analyse ultérieure. Répondez non. Ceux qui souhaitent plus de détails sur BAPS et ses différentes possibilités et menus sont invités à consulter la documentation livrée avec le logiciel.

Il s'agit ensuite de répéter le processus avec chacun des sous-échantillons. Ensuite, on charge le jeu de données brutes afin de le modifier. N'oubliez pas de créer un nouveau fichier de sortie à chaque fois. Dans chaque fichier de résultat BAPS sont donnés les clusters avec les individus qu'ils contiennent. Ces individus sont identifiés par leur rang d'entrée dans le jeu de données (1,2,3...). Par exemple, pour les femelles de Berne 1995, le fichier de résultat donne (en début de fichier) :

```
RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Data □le: IRTotBrut1Ber95F.gen
Number of clustered individuals: 22
Number of groups in optimal partition: 12
Log(marginal likelihood) of optimal partition: -384.965
Best Partition:
Cluster 1: {1}
Cluster 2: {2, 5, 16}
Cluster 3: {3, 20}
```

```
Cluster 4: {4}
Cluster 5: {6}
Cluster 6: {9, 15, 22}
Cluster 7: {8}
Cluster 8: {7}
Cluster 9: {12, 13, 17}
Cluster 10: {18}
Cluster 11: {10, 19}
Cluster 12: {11, 14, 21}
```

Le nom du fichier analysé est suivi de l'effectif de l'échantillon, du nombre de clusters dans la meilleure partition et de la valeur du Log de la valeur marginale de vraisemblance ou Log(MV) qui sert de critère à BAPS pour sélectionner la meilleure partition, c'est-à-dire celle qui présente le plus petit Log(MV). C'est bon à savoir si on souhaite relancer BAPS sur les mêmes données afin de voir s'il trouve une partition meilleure au deuxième essai. Enfin, la partition est donnée. Dans le jeu de données, il faut donc maintenant ajouter une colonne avec le numéro de cluster BAPS auquel chaque individu appartient. Il faut le faire pour tous les sous-échantillons (cf. tabl. 10). Attention, vous allez peut-être trouver des partitions légèrement différentes des miennes et avec des labels de clusters différents, c'est normal.

Il faut ensuite créer un nouveau fichier de données où chaque sous-échantillon initial se retrouve subdivisé en autant de sous-échantillons que de clusters de BAPS qui le composent (12 pour les femelles de Berne 1995). Sous un éditeur quelconque vous fusionnez les colonnes 1, 2, 3 et 6 du tableau 10, ce qui donne pour la première ligne quelque chose du style Ber95F1. N'oubliez pas de trier les données pour que les clusters apparaissent dans l'ordre dans chaque sous-échantillon initial. Appelons le fichier contenant ces données modifiées "IRTotBAPSClustMalHomoMFSep.txt".

Ce n'est pas fini, car il faut maintenant coder en données manquantes le locus IR08 chez les tiques mâles. Rappelez-vous que, dans un souci de puissance, nous les avions artificiellement rendus homozygotes afin que les mâles soient pris en compte pour ce locus dans l'analyse BAPS. Maintenant, nous souhaitons calculer les nouveaux F_{IS} de cette partition afin de voir si elle chute par rapport au jeu de données initiales. Le génotype des mâles au locus IR08 doit donc en effet être recodé 000000, car ils ne doivent pas rentrer en ligne de compte dans le calcul du F_{IS}. Pour ce faire, il est commode soit de faire un petit programme (pour ceux qui savent), soit d'utiliser la fonction conditionnelle d'Excel. Il s'agit de créer une colonne sexe en A dans le jeu de données "IRTotBAPSClustMalHomoMFSep.txt" avec le sexe des individus (F ou M), dans une colonne libre (en H après IR39) on tape en ligne 2 (ligne du premier individu) :

`SI(A2="M";"000000";C2)`, ce qui aura pour effet d'écrire "000000" dans la case H2 si l'individu est mâle ou de recopier le génotype de la femelle au locus IR08 (contenu dans la case C2). On copie ensuite H2 et on le colle de H3 à H726 (normalement

Tableau 10
Aspect du tableau de données brutes modifiées avec l'appartenance des individus aux clusters BAPS.

Site	An	Sexe	Individu	IndRang	Cluster BAPS	IR08	IR25	IR27	IR32	IR39
Ber	95	F	Bern95F_005	1	1	170183	150150	123123	235235	129129
Ber	95	F	Bern95F_007	2	2	174174	137146	119119	233250	133133
Ber	95	F	Bern95F_011	3	3	177183	000000	119119	243243	000000
Ber	95	F	Bern95F_013	4	4	173175	136142	119119	250250	142142
Ber	95	F	Bern95F_018	5	2	165178	137146	119119	243248	142142
Ber	95	F	Bern95F_020	6	5	165173	145148	119119	241241	129133
Ber	95	F	Bern95F_022	7	8	168171	134134	119119	243248	135135
Ber	95	F	Bern95F_027	8	7	171175	147147	119119	233233	125125
Ber	95	F	Bern95F_028	9	6	169175	140145	119119	233233	135142
Ber	95	F	Bern95F_029	10	11	166176	128145	119119	243243	125142
Ber	95	F	Bern95F_032	11	12	173183	134134	121121	233233	131137
Ber	95	F	Bern95F_037	12	9	175183	147147	119119	235235	134137
Ber	95	F	Bern95F_038	13	9	175183	135147	123123	250250	127127
Ber	95	F	Bern95F_039	14	12	183183	134134	119119	233243	121128
Ber	95	F	Bern95F_040	15	6	168174	141147	119119	233233	135142
Ber	95	F	Bern95F_042	16	2	174178	146146	119119	000000	112129
Ber	95	F	Bern95F_043	17	9	175175	000000	123123	233235	127134
Ber	95	F	Bern95F_044	18	10	174176	130130	119119	233233	128128
Ber	95	F	Bern95F_045	19	11	171175	145145	119121	243246	142142
Ber	95	F	Bern95F_048	20	3	173183	147147	119119	243243	129142
Ber	95	F	Bern95F_049	21	12	168170	000000	119121	233233	131144
Ber	95	F	Bern95F_050	22	6	169169	150151	119119	233233	129135
Ber	95	M	Bern95M_006	1	7	177177	134147	119119	233233	129129
Ber	95	M	Bern95M_008	2	8	172172	137148	119119	000000	000000
Ber	95	M	Bern95M_009	3	14	165165	146148	119127	248248	131137
Ber	95	M	Bern95M_010	4	3	000000	148148	123123	233233	131133

Figure 30
Importer les données dans Genetix.

la fin du fichier). On sélectionne les cases H2 à H726, on les copie et on fait un collage spécial (on veut ne coller que la valeur et non la formule) sur C2. On supprime les colonnes H et A et on sauve en texte seul sous le nom "IRTotBAPSClustMalManqIR08MFSep.txt". Supprimez aussi le label de la première colonne (c'est pour Genetix qui ne désire que le nom des loci).

Nous allons maintenant recalculer les F_{IS} par locus et sur l'ensemble, avec intervalles de confiance. Nous sommes paresseux et pour ne pas avoir à supprimer les clusters d'un individu pour lequel le calcul ne se fera pas, et étant donné que Fstat ne prend pas plus que 200 sous-échantillons (avec mes partitions je me retrouve avec 368 sous-échantillons), nous allons importer notre nouveau fichier sous Genetix. Lancez Genetix et allez dans le menu "Fichier", sélectionnez "Importer" et sélectionnez "Texte avec séparateur" et sélectionnez le fichier. Un menu apparaît et si vous avez fait comme moi, vous devez cocher les cases comme dans la figure 30. Quand cela est fait, cliquez "OK". Si le fichier est correctement chargé, cliquez dans le menu "Fstats" et sélectionnez "Weir & Cockerham". Cliquez OK dans la nouvelle fenêtre si vous ne changez pas le nom du fichier de sortie proposé "IRTotBAPSClustMalManqIR08MFSep.res". Après il faut prendre les résultats dans ce fichier en prenant garde que ce qui est annoncé comme écart-type des jackknives

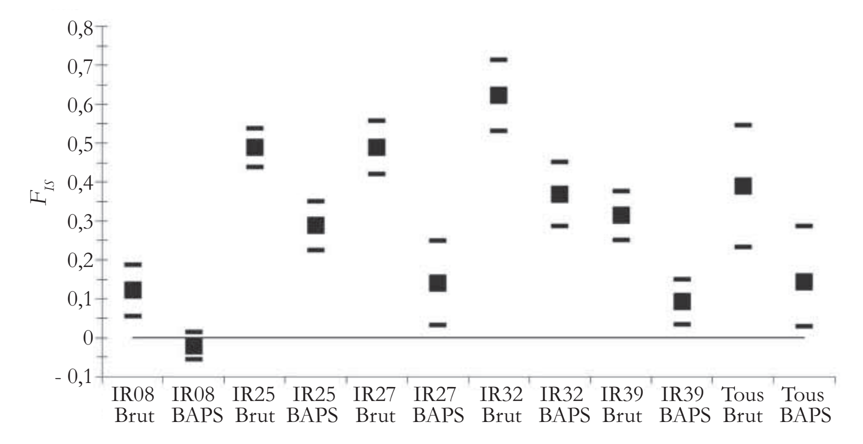


Figure 31
Comparaison de F_{IS} avant (données initiales : Brut) et après clusterisation par BAPS sur l'ensemble des données, par locus et sur l'ensemble (Tous).

correspond à l'erreur standard de Fstat. Il s'agit de comparer maintenant les F_{IS} de chaque loci et leurs intervalles de confiance de jackknife sur populations (voir p. 70-74 en partie 1) avant et après clusterisation par BAPS, ainsi que les valeurs globales et leur intervalle de confiance de bootstrap sur les loci (voir p. 70-74 en partie 1). La compilation des résultats prend alors la forme de ce qui est représenté dans la figure 31. Dans cette figure, il est aisé de voir que les clusters de BAPS présentent des déficits en hétérozygotes significativement inférieurs au F_{IS} de départ. Un test de rang de Wilcoxon pour données appariées confirme cela. Pour effectuer ce test sous R, il faut construire un fichier avec une colonne "Delta" où chaque ligne correspond à un locus.

Chaque valeur représente la différence entre le F_{IS} brut et le F_{IS} BAPS au locus correspondant (ici cinq valeurs). Appelons ce fichier "DeltaFisBrutBAPS.txt". Ensuite, sous R les commandes sont les suivantes :

```
> data<-read.table("DeltaFisBrutBAPS.txt",header=TRUE)
> attach(data)
> wilcox.test(Delta, alternative="greater")
```

Le test est unilatéral, car ce que nous recherchons est bien un effet Wahlund. Nous attendons au départ une chute du F_{IS}, d'où l'instruction "greater". La *P*-value = 0,031 obtenue est significative. Notons aussi que la plupart des loci, mis à part IR08, gardent un fort F_{IS} qui provient probablement des allèles nuls et autre dominance des allèles courts. Ces déficits restent très significativement au-dessus de 0 (fig. 31), ce qui rend bien compte du fait que "Hardy-Weinberg" n'est pas une

nécessité pour parvenir à une partition. Par ailleurs, le F_{IS} fait mieux qu'être faible pour IR08, il est négatif, ce qui est effectivement ce que nous attendons chez une espèce dioïque pangamique.

Il semble donc bien y avoir un effet Wahlund, contrairement à ce que la variance du F_{IS} entre loci pouvait laisser prévoir. Reste à déterminer si cet effet provient d'une micro-structuration (en groupes familiaux, par exemple) ou de la présence d'espèces (ou races d'hôtes, ou groupes adaptatifs ou écotypes) cryptiques. Afin d'essayer d'argumenter dans un sens ou l'autre, on peut essayer de regarder l'organisation de ces différents clusters. En principe, si on a à faire à différentes espèces, ces dernières devraient apparaître clairement. Si on effectue un arbre à partir d'une matrice de distance inter-clusters, ces derniers devraient être regroupés selon l'espèce à laquelle ils appartiennent en groupes séparés par des branches relativement longues comparées aux branches séparant chaque cluster (géographique, en principe) à l'intérieur de chaque espèce. Selon Takezaki et Nei (1996), la méthode du Neighbor-Joining (NJTree) sur distances harmoniques de Cavalli-Sforza et Edwards (1967) est une bonne solution. La matrice est obtenue en important "IRTotBAPSClustMalHomoMFSep.txt" dans Genetix[10], en cliquant sur le menu "Distances" puis "Sur données réelles" et en sélectionnant "Cavalli-Sforza & Edwards". On copie la matrice obtenue afin de l'incorporer dans un fichier de type MEGA (Kumar *et al.*, 2004) pour matrice de distances (ouvrir le fichier "IRTotBAPSClustMalHomoForNJTREENmini3CSE.meg" avec un éditeur de texte pour voir un exemple). Afin de limiter le nombre de branches et le poids des clusters ne contenant qu'un seul ou deux individus, je n'ai gardé que les clusters d'au moins 3 individus. L'arbre obtenu n'en est pas plus lisible pour autant et ce qui en ressort, c'est que les plus longues branches sont toujours celles séparant les clusters sans que se dégage une quelconque hiérarchie (on parle de râteau). Ceci plaide davantage en faveur d'une micro-structuration locale forte avec une différenciation géographique faible. On peut alors recommencer l'ensemble des opérations (BAPS->Genetix->MEGA) sur les échantillons de 1996 seuls et en ne séparant pas les mâles des femelles. Sur l'arbre obtenu, on ne voit pas mieux une quelconque structure sauf que les clusters tunisiens de plus de deux individus se retrouvent bien ensemble (fig. 32) avec un cluster du Tessin (Cen16 qui comporte d'ailleurs deux mâles et une femelle). Ceci ne contredit pas que l'effet Wahlund pourrait être issu de la présence dans chaque site d'individus issus des mêmes pontes. Ceci implique une forte variance du succès de survie entre pontes : de nombreux individus issus seulement de quelques pontes accèdent à l'âge adulte (voir Chevillon *et al.*, 2007a, pour un résultat similaire sur la tique du bétail).

[10] Je me suis rendu compte sur le tard que Genetix contenait quelques bugs dans ce module et je conseillerai d'utiliser plutôt MSA pour le calcul de distances, bien qu'ici cela n'ait pas changé grand-chose, raison pour laquelle j'ai laissé l'analyse telle qu'elle. Pour l'utilisation de MSA, se référer à la seconde partie de ce manuel, p. 254.

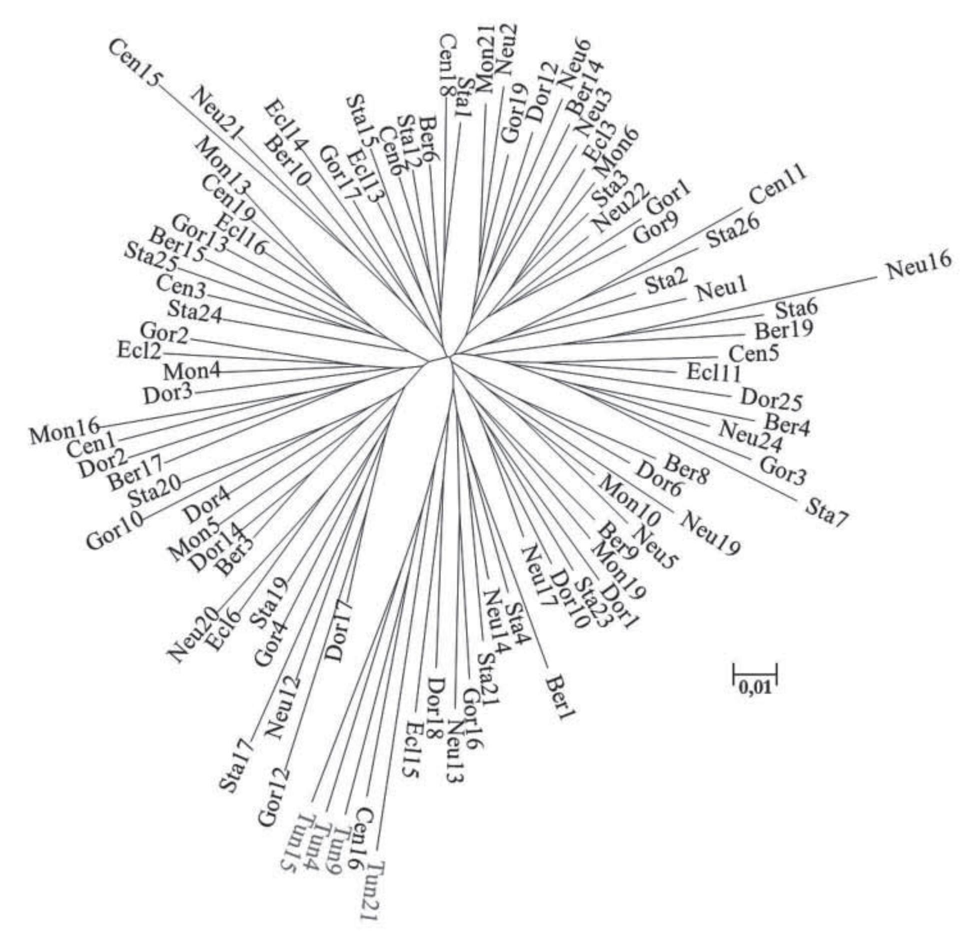


Figure 32
Dendrogramme exécuté selon la méthode du NJTREE sur les distances harmoniques de Cavalli-Sforza et Edwards entre paires de clusters BAPS de taille supérieure ou égale à 3 dans chaque sous-échantillon des tiques de 1996. Les clusters tunisiens sont indiqués en rouge.

Commentaires sur l’analyse des fichiers par BAPS

Contrairement à ce qui pourrait être suggéré à la lecture du manuel d’utilisation de BAPS, les clusters obtenus ne présentent pas ici une structure panmictique, mais conservent un déficit important d’hétérozygotes sauf pour IR08. Nous verrons, avec les analyses suivantes, que ces clusters reflètent probablement en grande partie une réalité biologique de nature assez complexe (races d’hôte, structures familiales), et qui devra conduire à d’autres études. BAPS ne

représente ici qu'un outil d'argumentation et d'orientation de futures investigations, pas un générateur de vérités.

CONCLUSION SUR LES DÉFICITS EN HÉTÉROZYGOTES

À l'occasion de ces premières analyses, nous pouvons constater qu'une analyse d'un jeu de données de génétique des populations requiert de la patience, de la méthode, ainsi qu'une bonne batterie de tests. Il était cependant nécessaire d'aller jusqu'au bout avant d'aller plus loin. Nous savons maintenant que ces tiques sont structurées à une échelle locale, ce qui explique une grande partie des déficits en hétérozygotes. Cet effet Wahlund résulte probablement d'une structure en groupes familiaux. L'existence d'espèces cryptiques n'est en effet pas soutenue par nos analyses NJTREE ni par l'absence totale de déséquilibre de liaison. Nous savons également qu'une partie non négligeable de ces déficits provient de l'existence d'allèles nuls (Loci IR25, IR32, IR39). Pour ces derniers, il y a donc un risque de surestimer la différenciation entre sous-échantillons, mais seulement pour des niveaux de différenciation atteignant au moins 10 % (F_{ST} = 0,1), en dessous de quoi l'effet devient faible (Chapuis et Estoup, 2007). Nous verrons que les niveaux de différenciation entre populations d'*I. ricinus* se trouvent bien en dessous de cette frontière. Enfin, un locus (IR27) a montré des évidences de dominance des allèles courts. Dans la mesure où ce phénomène modifie l'hétérozygotie et l'estimation des fréquences alléliques, il faudra être constamment vigilant quant aux résultats obtenus par la suite. Si nous avions un locus de plus sans allèle nul, j'aurais même conseillé de le supprimer. Ce n'est malheureusement pas le cas. Il faudra juste vérifier que chaque résultat ultérieur n'est pas sous la dépendance de ce seul locus. L'idéal aurait été d'avoir sept loci comme IR08, mais non liés à l'X ! Mais on ne choisit pas et les problèmes de marqueurs chez les parasites et vecteurs représentent un souci récurrent.

Une autre conclusion importante est qu'un déficit en hétérozygotes non entièrement expliqué par des allèles nuls exclut les causes endogamiques (croisements frère/sœur, autofécondation…) qui tendent à augmenter l'homozygotie et donc à dévoiler les homozygotes nuls (blancs), d'une part, et suggère, d'autre part, plutôt un effet Wahlund, qui augmente la diversité génétique H_s sans augmenter l'hétérozygotie observée (d'où augmentation du F_{IS}, cf. équation 19 en première partie de ce manuel, p. 47). Dans le cas d'un effet Wahlund, il est donc normal que les procédures de détection d'allèles nuls ne suffisent pas à expliquer entièrement les déficits en hétérozygotes, même si ces derniers sont présents, comme l'attestent la présence fréquente d'individus blancs, ainsi que la forte variance du F_{IS} entre loci.

STRUCTURE DES POPULATIONS ET SCHÉMAS DE DIFFÉRENCIATION

Nous avons ici une espèce à sexes séparés. La première chose à tester est s'il n'existe pas une différence entre femelles et mâles tiques, liée par exemple à un biais de dispersion spécifique de chaque sexe (Goudet *et al.*, 2002 ; Prugnolle et De Meeûs, 2002). En plus, on sait que c'est probablement le cas ici puisque ce signal fut détecté précédemment (De Meeûs *et al.*, 2002a), mais aussi lors de notre recherche de dominance des allèles courts. Par ailleurs, il est intéressant de vérifier si le signal persiste en tenant compte de la microstructure en clusters, même s'il a été montré que celle-ci a peu (pas) d'effet sur la structure à plus large échelle, si la microstructure n'est pas trop forte (Fontanillas *et al.*, 2004).

Structure génétique spécifique à chaque sexe des données brutes (sans tenir compte de BAPS)

Comment suspecter qu'un biais de structuration existe entre mâles et femelles ? Soit en effectuant directement le test "Sex biased dispersal" de Fstat, soit, comme cela a été le cas pour les données présentes, en testant la différenciation locale entre tiques mâles et femelles. La justification de ce test est qu'un tel signal avait été suggéré chez cette espèce en Irlande pour un locus enzymatique (Healy, 1979). Nous allons donc mesurer et tester la différenciation entre mâles et femelles de chaque sous-échantillon. Pour ce faire, il faut construire un fichier Fstat (par exemple) où femelles et mâles de chaque site sont considérés comme appartenant à des échantillons différents. Appelons ce fichier "IRTotBrutSexBias.dat" et chargeons-le dans Fstat (après avoir ouvert Fstat il faut aller dans "File", "Open", etc.). On coche ensuite dans Fstat l'option "Fst per pair of samples" et la case "Pairwise tests of differentiation", ainsi que la case "5/100" du "Nominal level for multiple tests". Cette dernière case est choisie pour gagner du temps. Ici, Fstat donnera le seuil Bonferroni de significativité corrigé par le nombre de tests (276 ici). À ce seuil, une *P*-value sera significative si elle est inférieure ou égale à 0,05/276 = 0,00018 et Fstat ajuste le nombre de permutations nécessaires pour atteindre cette valeur, soit 5 520, ce qui est bien suffisant. Avec "1/100" on obtient 27 600, ce qui est beaucoup. En plus, à ce niveau, le Bonferroni est beaucoup trop conservateur. De toutes façons, comme nous n'allons utiliser que les résultats par paire locale de femelles et de mâles, nous n'appliquerons pas cette procédure. Après avoir cliqué sur "Run" et attendu la fin des permutations, deux fichiers sont à consulter. "IRTotBrutSexBias.fst" donne les F_{ST} par paire et "IRTotBrutSexBias-pp.pvl" donne les *P*-value du test de randomisation des génotypes par paire de sous-échantillons. Dans ces fichiers, il faut garder les valeurs

correspondant aux paires femelle-mâle de chaque site-année. Si vous faites cela, deux probabilités sont significatives sur les 12 tests (17 %). Un test binomial peut alors être effectué sous R avec la commande suivante :

```
binom.test(2, 12, p=0.05, alternative="greater")
```

Le test est unilatéral, car on regarde si le nombre 2 n'est pas plus grand qu'attendu sous l'hypothèse nulle. Le test donne une *P*-value de 0,12, ce qui n'est pas vraiment significatif, mais témoigne d'un signal possible. Le test généralisé entrepris avec MultiTest et un *k*' = 6 donne une *P*-value globale seuil de 0,6015 (La notice d'utilisation de ce programme est suffisamment détaillée pour ne pas avoir à reproduire ici un tuteurage pas à pas).

Pour effectuer le véritable test de structuration sexe-spécifique, il faut remanier quelque peu le fichier initial des données afin de le mettre au format requis par Fstat pour l'analyse du biais de dispersion sexe-spécifique (Biased dispersal menu). Référez-vous à l'aide en ligne de Fstat pour construire ce fichier. Nous allons nous focaliser sur les échantillons 1996 uniquement. Une fois ce fichier constitué, il faut lancer Fstat, cliquer sur le menu "Biased dispersal" et y charger le fichier requis. Il faut ensuite sélectionner le test "Two sided" (on n'a en principe pas de préjugé pour l'instant) et cocher toutes les options comme dans la figure 33.

Vous remarquerez dans la figure 33 que les cases du F_{IS} et du H_o sont cochées comme les autres, alors que cela n'a aucun sens. En effet, puisque nous avons codé les mâles

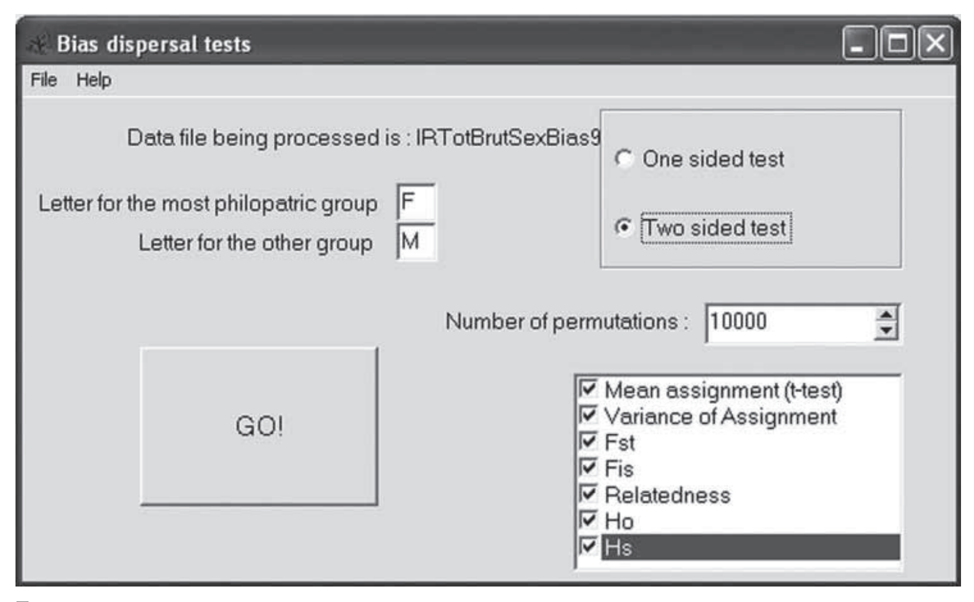


Figure 33
Menu et cases à cocher dans le menu "Biased dispersal".

homozygotes au locus IR08, il y aura nécessairement une différence mâle femelle à ce niveau. Cependant, quand cette option n'est pas cochée, on perd une partie de l'information sur H_s dans le fichier de sortie. Il conviendra donc, dans ce fichier, d'ignorer les résultats sur H_o et F_{IS}. Le logiciel crée cinq fichiers, trois fichiers .dat au format Fstat (les données totales, les femelles, les mâles), le fichier de permutations et le fichier .res des résultats (le plus utile). Ces derniers indiquent que les femelles sont bien mieux assignées que les mâles (IA_c = 0,36 et IA_c = - 0,56 pour les femelles et les mâles respectivement, *P*-value = 0,0005) et que les femelles sont localement mois diverses génétiquement (H_s = 0,79) que les mâles (H_s = 0,81) (*P*-value = 0,027), ce qui va dans le sens d'un biais de dispersion femelle (les femelles disperseraient moins). Par contre, le F_{ST} et la variance d'assignement répondent en sens inverse (mais non significativement heureusement). Pourtant, ce sont ces derniers paramètres (F_{ST} et variance d'assignement) qui doivent théoriquement signaler les premiers un biais de dispersion (qui donnent les tests les plus puissants) (Goudet *et al.*, 2002). Nous discuterons de ce paradoxe plus loin.

Afin de tester si la Tunisie n'est pas responsable seule de ce résultat, recommençons avec les données de Suisse 1996. Dans ce cas, on a des résultats comparables avec une *P*-value = 0,0004 pour l'assignement, mais une *P*-value = 0,06 marginalement significative pour H_s. Cantonnons-nous (normal pour la Suisse) au Plateau Suisse en excluant le site Monte-Ceneri du Tessin. Cette fois, les *P*-values tombent à 0,0002 et 0,02 pour les assignements et H_s respectivement. En restreignant l'échantillonnage aux sites du nord-ouest de la Suisse (il faut supprimer les sites Gorges-du-Trient et Dorénaz), sans oublier de le signaler en en-tête du fichier de données (il n'y a plus que cinq sites), on obtient une confirmation de ce qui était observé (tabl. 11), mais sur une échelle plus réaliste quant aux interprétations biologiques (en fin de ce chapitre). Il semble donc bien y avoir un biais de dispersion femelle (ou à tout le moins

Tableau 11
Résultats du test de biais de dispersion spécifique à chaque sexe sur les cinq sites du nord-ouest de la Suisse. Excepté la variance d'assignement ($s^2(AI_c)$), tous les autres paramètres plaident en faveur d'un biais de dispersion mâle (les femelles dispersent moins), avec une *P*-value (tests bilatéraux) très significative pour AI_c et F_{IS} et significative pour H_s.
Pour le F_{IS}, le test (unilatéral) a été réalisé en supprimant le locus IR08.

Paramètres	Femelles	Mâles	*P*-values
AI_c	0,523	- 0,786	0,0002
$s^2(AI_c)$	9,970	8,611	0,3425
F_{ST}	0,001	- 0,000	0,7964
H_s	0,776	0,813	0,0224
F_{IS}	0,422	0,506	0,0081

un biais de structuration génétique en faveur de ces femelles). En retirant chaque locus un à un et en recommençant l'analyse (donc cinq traitements), vous pourrez vérifier qu'aucun locus n'est responsable à lui seul du signal. On constate même, pour les données sans IR08, que le F_{IS} est significativement supérieur chez les mâles (tabl. 11). On pourra ici se contenter de refaire ces analyse sur les échantillons du Nord-Ouest et en unilatéral pour compenser la perte de puissance. La question qui se pose ensuite est de savoir si tenir compte des résultats de BAPS (microstructuration) change cette conclusion. Pour ce faire, il faut réanalyser les données en tenant compte des clusters définis par BAPS.

Structure génétique spécifique à chaque sexe des données clusterisées par BAPS

Nous prendrons ici le fichier de données 1996 de Suisse uniquement et les clusters obtenus en ne séparant pas les mâles des femelles (évidemment). Il faudra prendre garde à ne garder que les clusters contenant au moins une femelle et un mâle, car sinon Fstat va planter (comme on dit). Nous allons dans un premier temps effectuer l'analyse sur tous les clusters de tous les sites. Le label "Pop" va donc se positionner entre chaque cluster. On peut faire le test en unilatéral, mais au vu des résultats vous verrez vite qu'il convient de repartir sur une base de tests bilatéraux. Les résultats sont en effet spectaculairement divergents des précédents (tabl. 12).

Tableau 12
Résultats du test de biais de dispersion spécifique de chaque sexe d'*Ixodes ricinus* dans les cinq sites du nord-ouest de Suisse en tenant compte des clusters obtenus par BAPS (en ne séparant pas les mâles des femelles) et contenant au moins une femelle et un mâle. Tous les paramètres plaident fortement en faveur d'un biais de dispersion mâle (ils dispersent moins), avec des *P*-values (tests bilatéraux) très significatives sauf pour $s^2(AI_c)$ et F_{IS}. Pour le F_{IS}, le test a été réalisé en supprimant le locus IR08.

Paramètres	Femelles	Mâles	*P*-values
AI_c	- 0,160	0,216	0,0067
$s^2(AI_c)$	1,222	0,691	0,0706
F_{ST}	0,219	0,338	0,0012
H_s	0,584	0,524	0,0142
F_{IS}	0,189	0,297	0,0544

Ce résultat, très déconcertant au premier abord, est sous très forte influence du locus IR08, bien que les autres loci répondent dans le même sens (sauf peut-être IR32). Comme il s'agit peut-être d'un phénomène local, nous allons refaire les mêmes analyses, mais dans chaque site de 1996 (y compris la Tunisie). Le résultat

des tests sur le F_{ST} figure dans le tableau 13. Le signal reste le même, mais semble disparaître sans le locus IR08. Il se pourrait que ce locus soit diagnostique de certains groupes de tiques. Pour vérifier cela, il faut reprendre le fichier initial de données et grouper les individus, dans chaque site, selon leur génotype au locus IR08. Ce faisant, on recalcule sur cette nouvelle partition le F_{IS} et le F_{ST} avec Genetix, ce qui donne 0,47 et 0,02 respectivement, alors qu'on attend un faible F_{IS} et un fort F_{ST}. IR08 n'est manifestement diagnostique de rien du tout et le fait qu'il donne les meilleurs résultats provient vraisemblablement de sa qualité (peu ou pas d'allèles nuls et très faible variance des différents estimateurs).

Tableau 13
Résultat des tests de biais de dispersion spécifique de chaque sexe sur F_{ST}, effectués dans chaque site, entre les clusters définis par BAPS et contenant au moins une femelle et un mâle. Le test global est obtenu par une procédure binomiale généralisée et les tests sans IR08 ont été effectués de façon unilatérale (les mâles dispersent moins). Utiliser le fichier d'aide de MultiTest V.1.2. pour une description pas à pas de la procédure à suivre pour combiner les neufs tests.

Sites	Cinq loci	Sans IR08
Bern	0,3250	0,2431
Monte Ceneri	0,0817	0,2827
Dorenaz	0,3199	0,3355
Eclepens	0,1306	0,2700
Gorges du Trient	0,0159	0,6392
Montmollin	0,2422	0,9079
Neuchâtel	0,0636	0,4665
Staadswald	0,0426	0,1809
Tunisie	0,1272	0,0795
Tous (Binomial)	0,0041	0,2251

Il y a donc manifestement un effet cluster que nous essayerons d'interpréter plus loin. Afin de vérifier quand même si notre biais de dispersion spécifique femelle existe toujours même en tenant compte de l'effet Wahlund présent au sein de chaque site, la solution qui nous reste consiste à ne garder qu'un seul représentant ou une femelle et un mâle par cluster dans chaque site (nord-ouest de la Suisse 1996). On prendra le premier des individus ayant le génotype le plus complet de chaque cluster afin de conserver le plus de puissance possible. Par exemple, si dans un cluster d'un site quelconque, il n'y a que des mâles on ne prend qu'un individu, si possible génotypé aux cinq loci. Même chose pour des clusters de femelles. Pour les clusters mixtes, on prend la première femelle la plus

complète et le premier mâle le plus complet. On obtient ainsi un jeu de données de cinq sites avec un nombre d'individus fortement réduit par site. C'est aussi la raison pour laquelle les tests seront unilatéraux (les femelles dispersent moins). Le résultat de cette analyse figure dans le tableau 14 où on retrouve bien le signal initial suggérant un biais de dispersion femelle, à la différence que tous les paramètres vont dans le bon sens, même si c'est toujours AI_c qui donne la seule *P*-value significative.

Tableau 14
Résultat du biais de structuration femelle (unilatéral) sur le jeu de données réduit à un individu ou deux (une femelle et un mâle) par cluster BAPS pour les cinq sites du nord-ouest de la Suisse. Cette fois-ci, tous les paramètres vont dans le même sens (les femelles dispersent moins). Pour le F_{IS}, le test a été réalisé sans le locus IR08.

Paramètres	F	M	*P*-value
AI_c	0,496	- 0,520	0,0097
$s^2(AI_c)$	6,377	9,350	0,3341
F_{ST}	- 0,008	- 0,016	0,1307
H_s	0,824	0,847	0,1221
F_{IS}	0,470	0,511	0,2220

Interpréter l'ensemble des résultats sur les biais de structuration

Il semble bien y avoir un biais de dispersion biaisé pour les femelles (elles disperseraient moins) à l'échelle du plateau Suisse (ou même de régions plus restreintes), mais le signal est brouillé par une micro-structuration qui existe localement. Le fait que dans chaque site, les clusters trouvés par BAPS contiennent des femelles beaucoup plus hétérogènes que les mâles à l'intérieur de chaque cluster, mais beaucoup moins différentes d'un cluster à l'autre peut être interprété de deux façons. La première suggérerait que le biais de dispersion spécifique à chaque sexe s'inverse à petite échelle, mais on ne voit pas bien comment. La seconde suppose que les clusters correspondent plus ou moins à des frères et sœurs issus d'une même ponte et que les femelles ont une réussite beaucoup plus homogène que les mâles. Ne parviendraient à l'âge adulte, selon cette hypothèse, que beaucoup de mâles par ponte, mais de peu de pontes, alors que les femelles représenteraient un échantillon plus aléatoire des pontes (moins de sœurs que de frères dans chaque site). Pour confirmer cette interprétation, une approche théorique de modélisation/simulation serait nécessaire, mais dépasserait alors le cadre ambitionné par cet ouvrage. Enfin, ces clusters pourraient correspondre à des cohortes différentes (chevauchement de générations), très différenciées (dérive forte) et cela surtout pour les mâles dont beaucoup viennent d'ailleurs. Ici aussi, une approche

théorique s'avérerait nécessaire. Il est cependant raisonnable d'imaginer que si les larves et les nymphes mâles sont plus souvent retrouvées sur des hôtes très dispersants, alors il y a de fortes chances que chacun de ces individus hôtes porte des mâles apparentés (surtout les larves). Une fois dispersé et gorgé, chaque groupe a une chance très inégale de trouver un habitat favorable à la mue suivante. Il en résulterait que seuls certains groupes, parfois composés d'individus très apparentés (frères), survivraient dans une zone éloignée de leur site d'éclosion, alors que beaucoup de groupes mâles seraient éliminés. Si les larves et nymphes femelles préfèrent, quant à elles, les hôtes peu dispersants (petits rongeurs), il est probable que la survie de ces femelles soit distribuée plus aléatoirement entre femelles de pontes différentes. Ceci pourrait au final expliquer notre effet Wahlund produit en majorité par les tiques mâles.

Différenciation globale et isolement par la distance

Plusieurs éléments nous incitent ici à manquer d'optimisme. Il y a en effet de nombreux allèles nuls, un effet Wahlund local, de la dominance d'allèles courts à un locus, sans parler d'autres problèmes mis en évidence lors d'études de pedigrees (De Meeûs *et al.*, 2004a). Si on ajoute à cela que manifestement un biais de dispersion spécifique à chaque sexe existe, supposant qu'un des deux sexes migre beaucoup (voir Goudet *et al.*, 2002) et donc qu'une faible structuration en résulte nécessairement, la probabilité de trouver une structuration génétique devient faible, et c'est un euphémisme. Nous allons quand même tenter notre chance, et ce pour plusieurs raisons. D'abord, parce que nous ne sommes pas arrivés jusqu'ici pour se mettre à bailler aux corneilles, ensuite parce que « c'est la nuit qu'il est beau de croire à la lumière » (Rostand, 1908).

Définir différents niveaux de subdivision pour l'analyse hiérarchique

Nous ne considérerons ici que les échantillons de 1996. Nous pouvons envisager, grâce à HierFstat (Goudet, 2005), n'importe quelle structure du moment que cette dernière reste hiérarchique. Nous allons donc dans un premier temps considérer (référez-vous au besoin à la figure 15) l'Europe-Afrique comme tout, suivi de la Tunisie *versus* la Suisse, puis le Tessin *versus* le nord des Alpes et enfin le groupe Gorges-du-Trient, Dorénaz contre le plateau Suisse (Eclepens, Montmollin, Neuchâtel, Staadswald, Bern). Référez-vous à De Meeûs et Goudet (2007) pour des détails sur la confection d'un fichier HierFstat.

Analyse hiérarchique sur données brutes (pas de cluster BAPS)

Il faut donc créer un fichier avec quatre (hiérarchie) plus cinq (loci) colonnes. La première colonne correspond donc au continent, Cont avec 1 l'Europe (= la Suisse, et alors ?) et 2 pour l'Afrique (Tunisie). La deuxième colonne (NrdWTessin) va coder pour l'appartenance aux cantons du nord et nord-ouest de la Suisse (1), pour celle du Tessin (2)

(Monte-Ceneri) ou la Tunisie (3) qui n'est pas plus subdivisée, mais doit être aussi codée dans cette colonne. La troisième colonne (NrdWNS) correspond à l'appartenance ou non au nord-ouest (1) ou au sud-ouest (Gorges-du-Trient, Dorénaz = 2) de la zone du nord des Alpes suisses. Le Tessin et la Tunisie étant codés 3 et 4 respectivement dans cette colonne. La quatrième colonne (Site) correspond aux sites eux-mêmes (1 à 9). Les cinquième à neuvième colonnes correspondent aux cinq loci, le premier, IR08, étant codé homozygote pour les tiques mâles. Appelons le fichier ainsi construit "IRTot96HierFstat.txt". L'analyse va se faire sous HierFstat 0.04-4 (Goudet, 2006, mis à jour de Goudet, 2005) comme décrit dans De Meeûs et Goudet (2007). N'oubliez pas de remplacer les données manquantes "000000" par "NA". Lancez le logiciel R. Chargez le package HierFstat (Menu "Package", "Chargez le package", "hierfstat"). Changez de répertoire pour travailler dans celui où le fichier de données "IRTot96HierFstat.txt" se trouve (Menu "Fichier", "Changer le répertoire courant"). Dans la console R, tapez la succession de commandes (chaque ligne correspond à une commande devant être suivie d'un retour charriot), en respectant les majuscules et minuscules (distinctes en langage R):

```
> data<-read.table("IRTot96HierFstat.txt", header=TRUE)
> attach(data)
> loci<-data.frame(IR08,IR25,IR27,IR32,IR39)
> levels<-data.frame(Cont,NrdWTessin,NrdWNS,Site)
> varcomp.glob(levels,loci)
```

Cette dernière commande produit le résultat suivant :

```
$loc
            [,1]          [,2]          [,3]          [,4]      [,5]      [,6]
IR08  0.01223796  0.0001573914 -2.260871e-03  0.0022890321 0.4342422 0.4906015
IR25  0.01069015 -0.0029660662  1.666085e-03  0.0021349532 0.4523394 0.4658385
IR27  0.29270494 -0.0015575541  3.185784e-05 -0.0003405896 0.2581954 0.2624521
IR32  0.17740753 -0.0165926500  1.063656e-02  0.0070371095 0.4268548 0.3006536
IR39 -0.01488133  0.0438594202 -1.195459e-04  0.0001627161 0.2574235 0.6343434

$overall
       Cont  NrdWTessin      NrdWNS        Site         Ind       Error
 0.478159253 0.022900541 0.009954088 0.011283221 1.829055277 2.153889149
$F
                 Cont  NrdWTessin      NrdWNS        Site       Ind
Total       0.1061340 0.111217077 0.113426523 0.115930989 0.5219148
Cont        0.0000000 0.005686634 0.008158420 0.010960256 0.4651490
NrdWTessin  0.0000000 0.000000000 0.002485923 0.005303783 0.4620901
NrdWNS      0.0000000 0.000000000 0.000000000 0.002824882 0.4607495
Site        0.0000000 0.000000000 0.000000000 0.000000000 0.4592219
```

Dont l'interprétation est la suivante :

F_{IS} = 0,459 (nous retrouvons ici un résultat ancien et sans valeur, car les mâles sont artificiellement homozygotes ici au locus IR08), $F_{\text{Site/NrdWNS}}$ = 0,0028, $F_{\text{NrdWNS/NrdWTessin}}$

= 0,0025, $F_{NrdWTessin/Cont}$ = 0,0057 et $F_{Cont/Total}$ = 0,106. Toutes ces valeurs de différenciation sont très faibles sauf pour la Suisse contre la Tunisie. Il faut tester ensuite la significativité de ces différentes partitions en commençant par la plus incluse, le site :

```
> test.within(loci, test=Site, within=NrdWNS, nperm=1000)
$p.val
[1] 0.311
```

On voit bien que le site (comme on le craignait) n'influence en rien la partition de l'information génétique. Nous allons donc supprimer ce facteur de la hiérarchie :

```
> levels<-data.frame(Cont,NrdWTessin,NrdWNS)
> varcomp.glob(levels,loci)
$loc
              [,1]         [,2]           [,3]       [,4]       [,5]
IR08    0.01232344  0.000808808  -1.444965e-03  0.4355876  0.4906015
IR25    0.01077746 -0.002368730   2.440097e-03  0.4535566  0.4658385
IR27    0.29269212 -0.001654562  -8.948516e-05  0.2579981  0.2624521
IR32    0.17763798 -0.014577719   1.316236e-02  0.4309008  0.3006536
IR39   -0.01487489  0.043906268  -6.184974e-05  0.2575165  0.6343434
$overall
       Cont  NrdWTessin      NrdWNS         Ind       Error
 0.47855610  0.02611407  0.01400616  1.83555962  2.15388915
$F
                     Cont  NrdWTessin      NrdWNS          Ind
Total           0.1061541  0.11194680 0.115053669    0.5222206
Cont            0.0000000  0.00648061 0.009956456    0.4654790
NrdWTessin      0.0000000  0.00000000 0.003498519    0.4619924
NrdWNS          0.0000000  0.00000000 0.000000000    0.4601036
> test.within(loci, test=NrdWNS, within=NrdWTessin, nperm=1000)
$p.val
[1] 0.121
```

Le facteur NrdWNS, séparant les sites Dorénaz-Gorges-du-Trient de l'ensemble des sites suisses du Nord-Ouest, ne semble pas influencer davantage la structure génétique des tiques. Si nous le supprimons à son tour, nous obtenons :

```
> levels<-data.frame(Cont,NrdWTessin)
> varcomp.glob(levels,loci)
$loc
              [,1]           [,2]       [,3]       [,4]
IR08    0.01229331  -0.0003464944  0.4351133  0.4906015
IR25    0.01083164  -0.0004024918  0.4543119  0.4658385
IR27    0.29269022  -0.0017259148  0.2579689  0.2624521
IR32    0.17789976  -0.0042513096  0.4354972  0.3006536
IR39   -0.01487632   0.0438573712  0.2574958  0.6343434
```

```
$overall
        Cont   NrdWTessin         Ind       Error
  0.47883861  0.03713116  1.84038709  2.15388915
$F
                    Cont  NrdWTessin         Ind
Total          0.1061668  0.11439947   0.5224453
Cont           0.0000000  0.00921047   0.4657228
NrdWTessin     0.0000000  0.00000000   0.4607561
> test.within(loci, test=NrdWTessin, within=Cont, nperm=1000)
$p.val
[1] 0.058
```

Si nous choisissons de garder le facteur NrdWTessin (marginalement significatif, P-value = 0,058) cela aboutit à :

```
> test.between(loci, rand.unit=NrdWTessin, test=Cont, nperm=1000)
$p.val
[1] 0.331
```

Si on élimine le facteur NrdWTessin, il faut alors repasser par Fstat. Il n'y a en effet plus que trois niveaux hiérarchiques avec deux sous-populations représentées par l'ensemble des tiques suisses, d'une part et par celles de Tunisie, d'autre part. On aboutit à un F_{ST} = 0,113 très significatif (P-value < 0,0001) entre les tiques de Suisse réunies en une seule population et la Tunisie.

Avec un H_s = 0,832, cela correspond à un $F_{ST}' = F_{ST}/F_{ST\max}$ = 0,113/(1 - 0,832) = 0,673, ce qui est relativement considérable et témoigne du peu de migration entre les deux pays. Par contre, à l'échelle de la Suisse, cette migration est forte et même si les Alpes apparaissent comme un facteur limitant, tout semble se passer comme si, génétiquement au moins, on avait à faire à une seule unité à cette échelle.

Qu'en est-il si nous tenons compte des clusters trouvés par BAPS ?

Analyse hiérarchique sur données clusterisées par BAPS

Nous allons donc utiliser le fichier de données précédent avec une colonne supplémentaire correspondant aux clusters trouvés avec BAPS. En suivant alors une procédure identique à celle décrite plus haut, nous pouvons constater que les facteurs ClusterBAPS ($F_{\text{Clust/Site}}$ = 0,3, P-value = 0,001) qui mesurent la partition génétique entre clusters d'un même site, et Continent ($F_{\text{Continent/Total}}$ = 0,11, P-value = 0,001) qui mesure la différenciation entre Suisse et Tunisie, constituent les deux seuls facteurs qui structurent les sous-échantillons de façon significative.

Si nous ne gardons qu'un mâle ou une femelle ou un individu par cluster, comme pour le tableau 14, le résultat de l'analyse par HierFstat ne montre plus aucune différenciation, à moins d'ignorer tous les facteurs sauf le continent (analyse par Fstat, F_{ST} = 0,09, P-value = 0,001).

Test d'isolement par la distance

Nous ne travaillerons ici que sur les échantillons suisses de 1996. D'abord parce que la Tunisie est trop éloignée par rapport aux distances entre échantillons suisses. Il y aurait deux groupes de points. Procéder à un test de régression entre deux points n'a pas de sens, le plus court chemin entre eux étant nécessairement une droite, c'est dans tous les bons livres de statistiques. Or, le test d'isolement par la distance est une forme de régression où on cherche à expliquer une différence génétique croissante par un éloignement géographique. Ensuite, il n'y a pas assez d'échantillons en 1995.

Pour le test, il faut configurer un fichier avec deux demi-matrices, l'une pour les distances géographiques entre paire de sites et l'autre pour les F_{ST} (estimés par θ) correspondants. Pour les distances géographiques, vous pouvez vous aider de la figure 15. Pour les F_{ST}, il suffit de prendre la sortie "IRTot96CH.fst" que Fstat a produit en analysant le fichier "IR96CH.dat" des données suisses 1996, si vous avez toutefois coché la case "Fst per pair of samples". En ce qui me concerne, j'obtiens les matrices représentées dans le tableau 15. Le test va être effectué selon la méthode décrite par Rousset (1997) pour un schéma en deux dimensions. Nous allons donc effectuer un test de Mantel sur la corrélation entre le $F_{ST}/(1 - F_{ST})$ et le log népérien (ou naturel) de la distance géographique. Nous allons utiliser Genepop 3 pour faire ce test et donc formater les données dans ce sens et les sauvegarder dans un fichier que nous appellerons IR96CH.mig. Ce fichier doit être configuré comme présenté dans la figure 34.

```
From·file:·IRCH96¶
8·pop¶
Intraclass·estimate·(Fwc_st):¶
0.0002¶
·0.0080··0.0012¶
-0.0003·-0.0049·0.0072¶
·0.0040·-0.0015·0.0049··0.0015¶
·0.0040··0.0085·0.0224··0.0078·0.0143¶
-0.0005·-0.0033·0.0042·-0.0015·0.0014·0.0059¶
·0.0116··0.0058·0.0136··0.0132·0.0042·0.0209·0.0089¶
distances:¶
·85.53¶
·50.00··46.05¶
·43.42··52.63···7.89¶
·19.74··65.79··26.32··19.74¶
105.26··78.95·102.63·102.63·102.63¶
115.79··82.89·110.53·111.84·113.16··11.84·¶
171.05·218.42·213.16·207.89·190.79·160.53·165.79¶
```

Figure 34
Présentation du fichier pour tester l'isolement par la distance entre sites de prélèvement d'*Ixodes ricinus* en Suisse en 1996 sous Genepop.

Tableau 15
Distances géographiques en km et différenciation génétique mesurée par le F_{ST} (Theta) par paire de sites d'échantillonnage d'*Ixodes ricinus* (abréviations comme dans la figure 15).

Theta							
Site	**Ber**	**Ecl**	**Mon**	**Neu**	**Sta**	**Dor**	**Gor**
Ecl	0,0002						
Mon	0,0080	0,0012					
Neu	- 0,0003	- 0,0049	0,0072				
Sta	0,0040	- 0,0015	0,0049	0,0015			
Dor	0,0040	0,0085	0,0224	0,0078	0,0143		
Gor	- 0,0005	- 0,0033	0,0042	- 0,0015	0,0014	0,0059	
Cen	0,0116	0,0058	0,0136	0,0132	0,0042	0,0209	0,0089
Distance en kilomètres							
Ecl	85,53						
Mon	50,00	46,05					
Neu	43,42	52,63	7,89				
Sta	19,74	65,79	26,32	19,74			
Dor	105,26	78,95	102,63	102,63	102,63		
Gor	115,79	82,89	110,53	111,84	113,16	11,84	
Cen	171,05	218,42	213,16	207,89	190,79	160,53	165,79

L'étape suivante consiste à lancer Genepop 3. Éviter de double cliquer sur le fichier genepop.bat, mais préférez ouvrir une session DOS en lançant une “Invite de commandes” dans le menu “Accessoires” de Windows. Dans la fenêtre DOS, et si Genepop est dans le répertoire “Genepop” du disque D, tapez “D:”, puis “Entrée”, puis “cd Genepop”, puis “Entrée”. Vous êtes dans le répertoire Genepop. Tapez alors “isolde”, puis “Entrée” pour lancer le programme d'isolement par la distance. À l'invite, tapez le nom complet du fichier de données puis “Entrée”. Le logiciel vous demande ensuite quel type de distance (non transformée ou Log) et quel type de mesure de différenciation vous souhaitez tester (X, qui figure dans la matrice ou X/(1 - X)). À vous de choisir la méthode appropriée. Genepop vous demande ensuite la distance minimale en deçà de laquelle la mesure de corrélation ne tient plus compte des données, car en deçà d'un certain niveau la réponse a en effet tendance à ne plus suivre un modèle clair (Rousset, 1997). Réfléchissez à ce que devrait

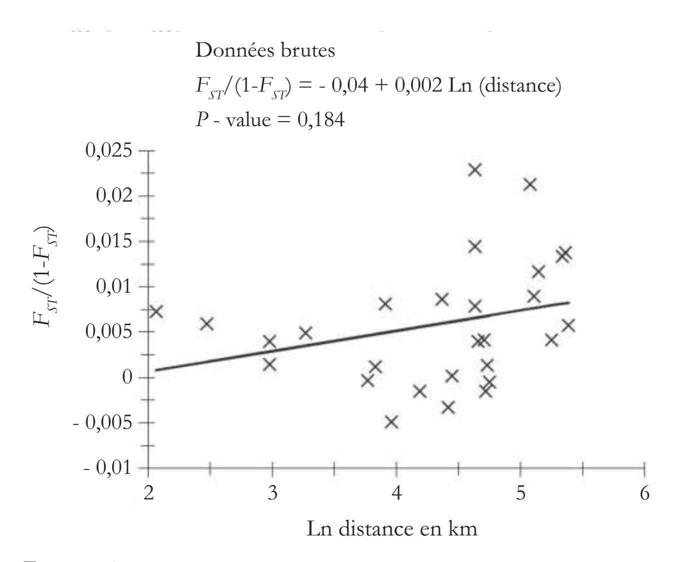


Figure 35
Représentation de l'isolement par la distance entre sites suisses pour les tiques récoltées en 1996. La distance minimale de 2 km a été choisie.

être cette distance minimale et tapez-la. Le nombre de randomisations vous est ensuite demandé. Tapez 1 000 000 pour être sûr d'obtenir une *P*-value suffisamment précise. Genepop vous demande, comme pour tous ses tests passant par randomisation, quatre nombres qui serviront de « graines » pour générer des nombres pseudo-aléatoires qui conditionnent le départ des randomisations. Tapez ce que vous voulez entre 1 et 168, comme indiqué avec un retour charriot après chaque chiffre. Quand les randomisations sont terminées, Genepop le signale avec un bip assez désagréable, mais qui ne doit pas vous effrayer (en général, je coupe le son avant). D'après une collègue avisée (TG), il n'y a pas de son sur la version Web du programme qui, par ailleurs, est sensiblement plus conviviale. Genepop a alors créé deux fichiers, l'un porte l'extension .ISO qui vous donne le résultat du test avec les paramètres de la régression et la *P*-value. Le second porte l'extension .GRA et donne les coordonnées en colonne de cette régression afin de pouvoir dessiner le graphique correspondant, comme représenté dans la figure 35. On y remarque que la relation n'est pas significative. Il semble cependant qu'une tendance existe. Peut-être l'existence d'une sous-structure nuit-elle à la clarté du signal ? Pour le vérifier, il suffit de procéder au même test, mais avec les données où un ou deux (de chaque sexe) individus par cluster avaient été gardés (voir p. 151-152). Le résultat change du tout au tout puisque la relation devient très significative, comme en témoigne la figure 36. Ceci permet de calculer le voisinage $Nb = 1/b = 173$ individus (WATTS *et al.*, 2007), le produit de la densité d'individus par km² par la surface de dispersion des descendants reproducteurs par rapport à leurs géniteurs, en utilisant la méthode de

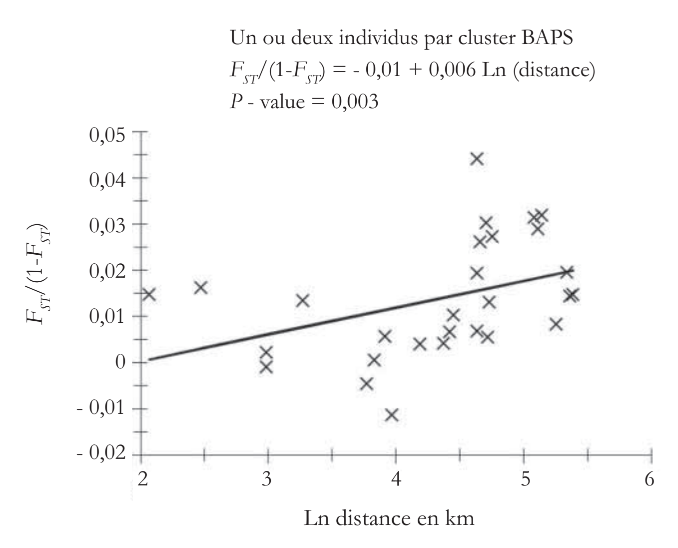


Figure 36
Représentation de l'isolement par la distance entre sites suisses pour les tiques récoltées en 1996 en ne gardant qu'un ou deux individus des clusters définis par BAPS dans chaque site. La distance minimale de 2 km a été choisie.

Rousset (1997) décrite en première partie (p. 88-90), ce qui donne $D\sigma^2 = 1/(4\pi 0{,}00577) = 13{,}78$. On peut aussi estimer le nombre d'immigrants présents dans une sous-population (Rousset, 1997), $Nm = 1/2\pi b = 28$ individus par génération. Il serait dommage de s'arrêter ici et nous allons donc essayer d'appréhender, même grossièrement, dans quelle gamme de valeurs se situe la densité de tiques afin d'en pouvoir extraire la surface de dispersion de ces tiques.

Estimation d'effectifs efficaces, extrapolation des densités et de la dispersion

Effectifs efficaces des tiques de Suisse

Pour des raisons de commodité et de cohérence (les tiques tunisiennes n'ayant pas été échantillonnées de la même manière), nous nous focaliserons sur les échantillons de Suisse. Rappelons-nous que nous avons rencontré de gros déficits en hétérozygotes (allèles nuls et dominance d'allèles courts), ainsi que la présence d'un fort effet Wahlund. Nous ne travaillerons donc qu'à partir de méthodes indépendantes de l'hétérozygotie, telle que celle proposée par Bartley *et al.*(1992), basée sur les déséquilibres de liaison et implémentée par NeEstimator (Peel *et al.*, 2004). Nous n'utiliserons que les données clusterisées par BAPS où seuls subsistent un ou deux individus par cluster dans chacun des huit sites suisses pour éviter l'effet confondant dû à l'effet Wahlund.

Il faut créer un fichier par site dans un format proche de Genepop comme dans la figure 37.

```
Pop¶
Bern95F_044,·174176·130130·119119·233233·128128¶
Bern95M_030,·174174·134134·119119·233241·134137¶
Bern95F_027,·171175·147147·119119·233233·125125¶
Bern95M_019,·171171·134147·119119·233233·131131¶
Bern95F_039,·183183·134134·119119·233243·121128¶
Bern95M_006,·177177·134147·119119·233233·129129¶
Bern95F_013,·173175·136142·119119·250250·142142¶
Bern95M_035,·172172·140147·119119·000000·129129¶
Bern95F_028,·169175·140145·119119·233233·135142¶
Bern95M_055,·168168·145151·119119·233233·135135¶
Bern95F_022,·168171·134134·119119·243248·135135¶
Bern95M_031,·169169·147147·119123·233233·130130¶
Bern95F_018,·165178·137146·119119·243248·142142¶
Bern95M_009,·165165·146148·119127·248248·131137¶
Bern95F_042,·174178·146146·119119·000000·112129¶
Bern95M_057,·174174·000000·119119·246246·121129¶
Bern95F_020,·165173·145148·119119·241241·129133¶
Bern95M_014,·173173·145145·119119·250250·129144¶
Bern95F_029,·166176·128145·119119·243243·125142¶
Bern95F_049,·168170·000000·119121·233233·131144¶
Bern95M_041,·170170·134148·119119·233233·142142¶
Bern95F_005,·170183·150150·123123·235235·129129¶
Bern95M_052,·170170·145145·119123·233233·127129¶
Bern95F_037,·175183·147147·119119·235235·134137¶
Bern95M_017,·175175·141141·119119·248248·133133¶
Bern95M_015,·175175·148148·119119·233241·126133¶
Bern95M_026,·192192·152152·119125·233233·149149¶
Bern95F_007,·174174·137146·119119·233250·133133¶
Bern95M_054,·172172·146146·119119·233233·125125¶
Bern95M_051,·169169·144144·123123·233243·136138¶
Bern95F_032,·173183·134134·121121·233233·131137¶
Bern95M_025,·176176·134144·121121·000000·000000
```

Figure 37
Format de fichier pour NeEstimator pour les tiques de Berne 1995.

Il faut ensuite lancer le programme NeEstimator (après l'avoir installé sur votre machine, bien entendu). Une fenêtre d'avertissement sur le copyright et sur la manière idoine de citer ce logiciel apparaît. Cliquez sur OK pour accéder au programme qui apparaît dans une fenêtre comme dans la figure 38. Comme indiqué sur la figure 38, cliquez sur le menu déroulant "File" et "Open", ce qui permet d'ouvrir la fenêtre "Analysis".

Dans la fenêtre "Analysis", une série d'onglets apparaît et vous positionne sur celui du format de vos données "Data Format" où il n'y a rien à changer, car vous avez



Figure 38
Menu à l'ouverture de NeEstimator.

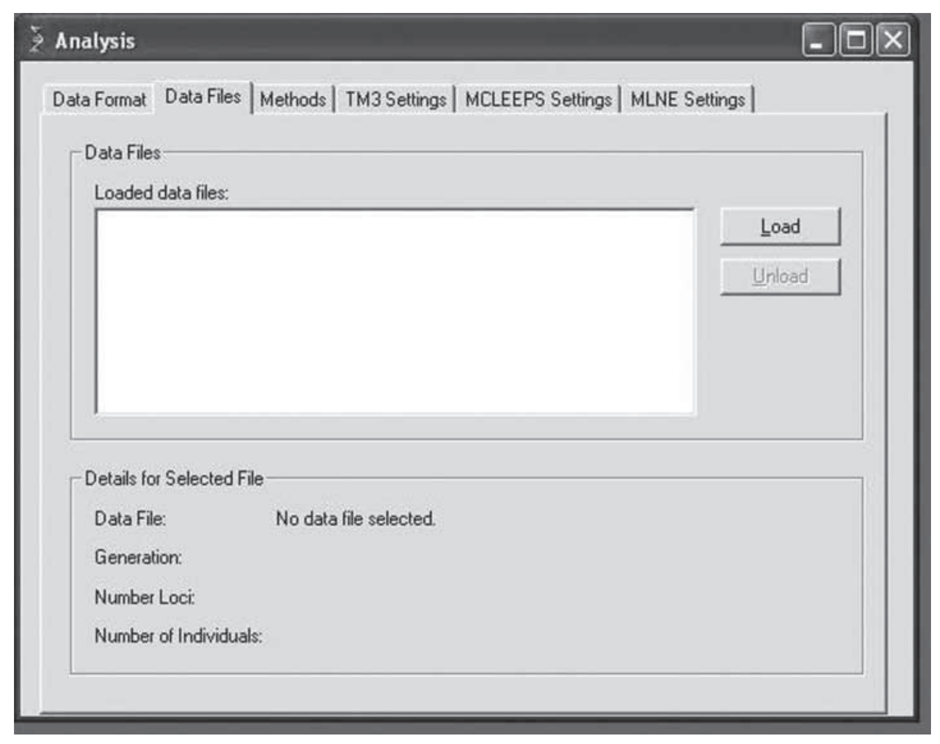


Figure 39
Onglet "Data File" avec le bouton "Load" qu'il faut clicker.

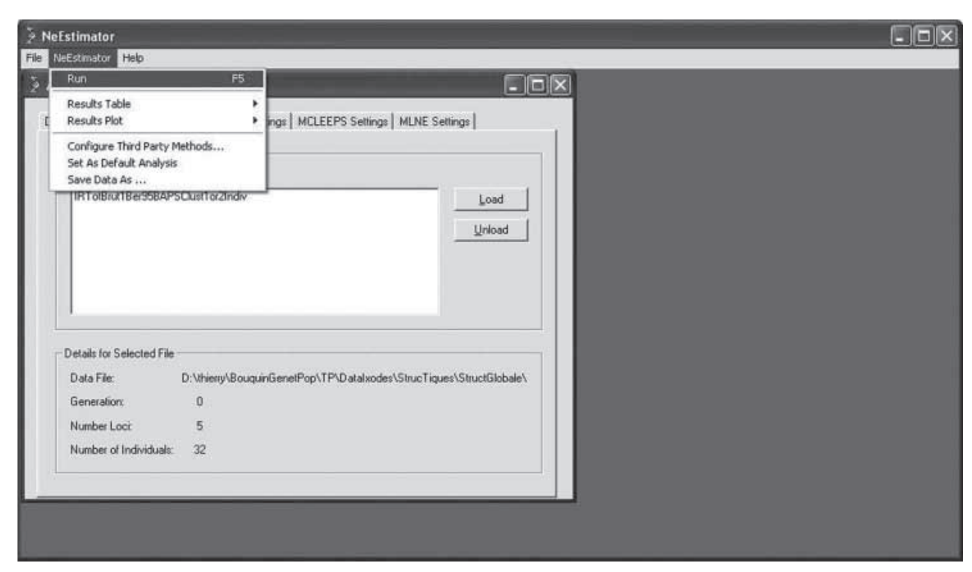


Figure 40
Pour lancer l'analyse de NeEstimator.

choisi le format par défaut. Allez à l'onglet "Data Files". Là il n'y a qu'un seul bouton "Load" qui vous permet de charger votre jeux de données, ce que vous faites (fig. 39). Une fois que vous avez choisi le fichier, le logiciel vous demande à quelle génération ces données correspondent-elles. Laissez la valeur par défaut "0", car nous n'utiliserons pas ici la méthode des moments de Waples (1989) (cf. p. 102 en première partie) et cliquez sur "OK". Dans le menu déroulant "NeEstimator", cliquez sur "Run" (fig. 40). Ce qui fait apparaître un message qui vous avertit qu'avec un seul échantillon, on ne peut utiliser les méthodes temporelles "Moment based" et vous demande si vous souhaitez continuer avec les méthodes à un seul échantillon. Vous répondez "Oui" bien entendu. Le résultat est affiché sous forme de tableau que je vous conseille de sauvegarder au format NeEstimator (NeA). Je conseille aussi de transcrire tous les résultats dans un tableur au fur et à mesure afin de disposer de l'ensemble dans un seul fichier. C'est ce qui est représenté dans le tableau 16.

Ici, bien que nous disposions d'échantillons espacés dans le temps (Bern, Gorges-du-Trient et Staadswald), ces échantillons ne sont séparés que d'une année, soit environ 1/3 du temps de génération d'*I. ricinus*. Ici, les adultes présents d'une année sur l'autre font partie de cohortes séparées et qui, même à long terme, auront du mal à échanger des gènes. La différenciation entre ces cohortes, déjà remarquée par De Meeûs *et al.* (2002a), va tendre à être très supérieure à celle qui existe réellement entre deux générations d'adultes reproducteurs. L'utilisation des méthodes temporelles sur nos données aboutira donc à de fortes sous-estimations des effectifs efficaces. Faites-le et vérifiez qu'effectivement, compte tenu qu'il n'y a qu'un tiers de

Tableau 16
Résultats synthétiques obtenus pour le calcul des effectifs efficaces (N_e) et leur intervalle de confiance à 95 % (*Li* et *Ls*) par la méthode des déséquilibres de liaison dans NeEstimator. Les valeurs infinies sont ignorées pour le calcul des moyennes. Les échantillons de 1995 sont considérés comme indépendants, car appartenant à des cohortes de tiques génétiquement isolées de celles de 1996 (le cycle d'*Ixodes ricinus* dure environ trois ans).

Échantillon	N_e	*Li*	*Ls*
Berne 1996	73	45	182
Berne 1995	222	79	Infini
Monte-Ceneri 1996	Infini	288	Infini
Dorénaz 1996	700	124	Infini
Eclépens 1996	Infini	81	Infini
Gorges-du-Trient 1995	177	10	601
Gorges-du-Trient 1996	75	43	219
Montmollin 1996	338	87	Infini
Neuchâtel 1996	398	93	Infini
Staadswald 1995	161	84	1 164
Staadswald 1996	Infini	374	Infini
Moyenne totale	268	119	541

génération séparant 1996 de 1995, les estimations obtenues par la méthode de Waples donnent des effectifs efficaces proches de 0, ce qui n'est pas très conforme à la perception que l'on peut avoir sur le terrain.

En reprenant le tableau 16, nous obtenons par conséquent un effectif efficace de 268 en moyenne sur l'ensemble des échantillons avec un intervalle de confiance à 95 % de [119, 541], avec des valeurs minimales et maximales de 73 et 700 respectivement. Ces nombres paraissent plausibles, compte tenu de l'effet Wahlund reflétant probablement un fonctionnement particulier des populations de tiques susceptible d'en réduire sensiblement l'estimation de leurs effectifs efficaces.

En reprenant les données avec un ou deux individus par cluster BAPS, les valeurs obtenues sont plus grandes en moyenne (596) avec un minimum et un maximum de 75 et 1 057 respectivement[11].

[11] Sur ces mêmes données, l'estimation avec un logiciel alternatif, LDNe (Waples et Do, 2008), non encore connu au moment de la rédaction de ce chapitre et dont l'utilisation est détaillée plus loin, donne une moyenne de $N_e = 223$.

Extrapolation des densités et des distances de dispersion des tiques en Suisse

Il faut dans un premier temps estimer sur quelle surface se distribuent les tiques. Ici, c'est difficile et on ne peut pas dire grand-chose de plus que les surfaces d'échantillonnage s'étendaient grossièrement sur S = 0,2 km². Ceci signifie (mais vous vous en doutiez probablement) que les estimations à venir seront tout à fait approximatives. À partir de là, les densités sont faciles à calculer (N_e/S). La densité moyenne devient 1 340 tiques reproductrices/km² 95 % CI = [594, 2 706] avec un minimum et un maximum de 367 et 3 502 tiques/km² respectivement (tabl. 16). En réutilisant les résultats de la régression de l'isolement par la distance $D_e\sigma^2$ = 13,78 (voir p. 160), on aboutit à une surface de dispersion moyenne entre adultes et leurs parents d'environ 0,01 km² [0,005, 0,023] avec un minimum et un maximum de 0,004 et 0,038 km² respectivement. Autrement dit, la distance moyenne séparant un adulte reproducteur de ses géniteurs est d'un ordre de grandeur de 100 m par génération (donc tous les trois ans environ), un intervalle de confiance à 95 % de bootstrap = [71, 152] et un maximum et un minimum de 63 à 195 m, ce qui est relativement modeste. Les données clusterisées par BAPS conduisent à une densité de 3 000 tiques par km² et une dispersion de moins de 60 m par génération. Donc, sachant que l'estimateur sans doute le moins biaisé est le produit $D_e\sigma^2$, la dispersion par génération est, quoi qu'il en soit, extrêmement modeste à moins d'évoquer des densités (effectifs) efficaces extrêmement faibles. Il en va donc de même en ce qui concerne la propagation des maladies par les tiques.

CONCLUSIONS SUR LA BIOLOGIE ET LA GÉNÉTIQUE DES POPULATIONS D'*I. RICINUS* EN SUISSE

Il existe un déficit important en hétérozygotes dans les populations d'*I. ricinus* (F_{IS} = 0,39) dont une majeure partie (64 %) est expliquée par un effet Wahlund important.

Le F_{IS} = 0,14 résiduel correspondrait à du « stuttering », à de la dominance d'allèles courts et à des allèles nuls. Pour tester les allèles nuls dans les clusters de BAPS, on ne peut pas utiliser Micro-Checker (échantillons trop petits). Nous pouvons néanmoins tester s'il existe une relation positive entre le nombre de blancs à un locus et le F_{IS} à ce locus. En effet, en reprenant les données clusterisées et en séparant les mâles des femelles en deux fichiers, il est facile de compter les blancs pour chaque locus avec la fonction "SI" d'Excel. Il suffit de créer

autant de nouvelles colonnes qu'il y a de loci et de remplir chacune avec les instructions de type "= SI(G2 = "000000") ; 1;0)" pour inscrire "1" quand on a un blanc. À la fin de chacune de ces colonnes, on tape une instruction du type "= somme(L2:L147)" pour obtenir la totalité des blancs à ce locus sur l'ensemble des clusters. Le F_{IS} de chaque locus est récupérable dans les deux fichiers de sortie Fstat de l'analyse des deux jeux de données (un pour le femelles et un pour les mâles) avec les données clusterisées par BAPS, que j'ai personnellement nommés IRTotBAPSClustMalManqIR08Females.dat et IRTotBAPSClustMalManq IR08FMales.dat respectivement, et où on aura pris soin d'éliminer le locus IR08 du fichier des mâles. Quand on a fait ceci pour les femelles et les mâles, on obtient le jeu de données présenté dans le tableau 17. La corrélation entre le nombre de blancs et le F_{IS} peut être analysée par un test de corrélation de Spearman (test non paramétrique). Ce test est facile à réaliser sous R. Si le fichier de données correspondant au tableau 17 s'appelle "AllelesNulsClustersBAPS.txt", alors il suffit de lancer R, et de se placer dans le répertoire contenant ce fichier (menu déroulant "Fichier", "Changer le répertoire courant").

Tableau 17
Données pour la régression entre le nombre de données manquantes (génotypes « blancs ») et la valeur des F_{IS} pour les différents loci (chez les mâles et les femelles pris séparément).

Sexe	Locus	Blancs	F_{IS}
Femelles	IR08	10	- 0,030
	IR25	50	0,256
	IR27	22	0,201
	IR32	47	0,253
	IR39	45	0,076
Mâles	IR25	51	0,368
	IR27	21	0,010
	IR32	74	0,473
	IR39	30	0,115

Ensuite, il faut taper les instructions suivantes :

```
> data<-read.table("AllelesNulsClustersBAPS.txt",header=TRUE)
> attach(data)
> cor.test(data$NBlancs, data$FIS, alternative="two.sided",
method="spearman")
```

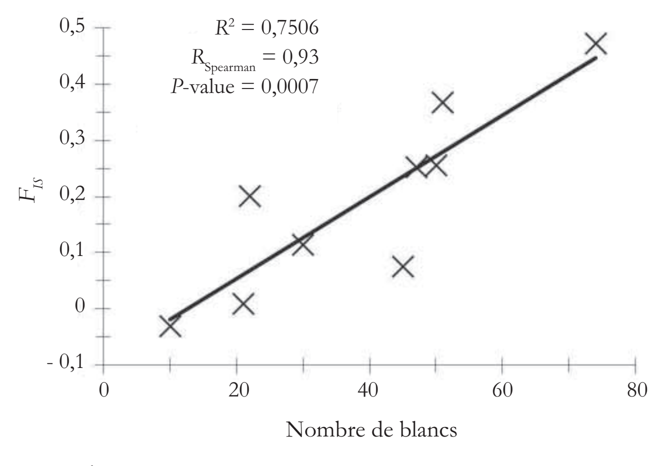


Figure 41
Relation entre le nombre de génotypes blancs trouvés par locus et le F_{IS} mesuré à ce locus sur l'ensemble des clusters de BAPS de l'ensemble des données microsatellites d'*Ixodes ricinus* (1995-1996, Tunisie et Suisse).

ce qui renvoie au résultat :

```
       Spearman□s rank correlation rho
data:  data$NBlancs and data$FIS
S = 8, p-value = 0.0007496
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.9333333
```

La figure 41 illustre la relation positive forte entre les deux variables.

Nous pouvons également tester de nouveau la dominance d'allèles courts au locus IR27 en prenant les F_{IS} par allèle donnés par la sortie Fstat sur les mâles et les femelles séparément. Nous ne distinguerons en effet pas les clusters, car il y en a trop. Sous R, la procédure est comparable à celle utilisée en p. 129-134, sauf que nous n'utiliserons que le polynôme d'ordre deux de la taille des allèles et le sexe comme variables explicatives du F_{IS}. Le résultat est de nouveau très significatif et on explique donc toujours une forte proportion du F_{IS} par ce phénomène de dominance des allèles courts, comme illustré par la figure 42. Suivent les instructions R :

```
> data<-read.table("FISAlleleSizeIR27ClustersBAPS.txt",header=TRUE)
> attach(data)
> loc27<-glm(data, formula = Fis ~ poly(Allele, 2) + Sexe, family = gaussian)
> anova(loc27, test="F")
```

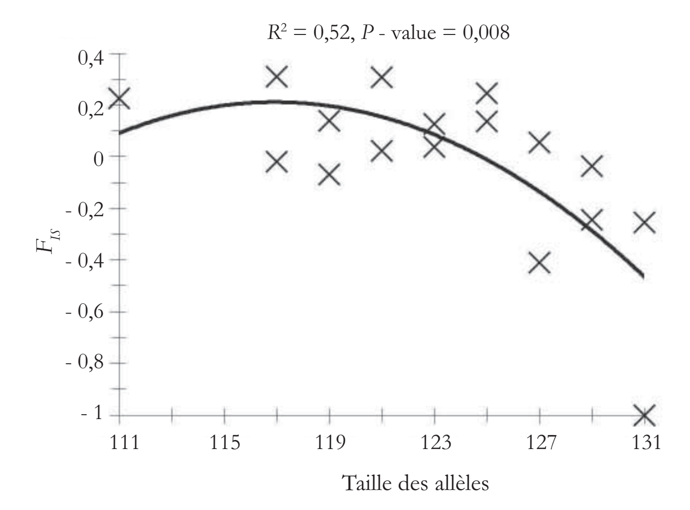


Figure 42
Régression entre taille des allèles et F_{IS} observés au locus IR27 dans les clusters d'*I. ricinus* définis par BAPS. Il n'y avait pas assez de données pour calculer des intervalles de confiance.

ce qui renvoie au résultat suivant :

```
Analysis of Deviance Table
Model: gaussian, link: identity
Response: Fis
Terms added sequentially (first to last)
                Df Deviance  Resid. Df  Resid. Dev     F       Pr(>F)
NULL                                16        1.65129
poly(Allele, 2) 2  0.85916          14        0.79212 7.0896  0.008281 **
Sexe            1  0.00441          13        0.78771 0.0728  0.791550
```

Comme cela a été vu au début de ce paragraphe, la majeure partie (64 %) du F_{IS} initial est expliquée par un effet Wahlund. Cet effet Wahlund est plus prononcé chez les mâles dont les clusters se trouvent plus différents entre eux que les femelles. Dans l'hypothèse de spécificités d'hôtes différentes des larves et/ou nymphes femelles et mâles, des groupes de larves ou nymphes mâles fortement apparentés seraient transportés ensemble sur le même hôte très dispersant (oiseau) avec de forts risques de tomber sur des sites défavorables lors du détachement, à la fin du repas sanguin. Les mâles retrouvés adultes dans nos échantillons correspondraient alors aux quelques groupes d'apparentés ayant eu la chance de tomber ensemble dans un site favorable. Les larves ou nymphes femelles seraient, quant à elles, plus souvent retrouvées sur des hôtes très peu dispersants, comme des petits rongeurs très territoriaux. Il en résulterait un apparentement réparti beaucoup plus aléatoirement pour les femelles dans chaque site. Il y a un fort biais de dispersion spécifique à chaque sexe (les

femelles dispersent très peu). Ce biais est partiellement masqué par l'effet Wahlund, et il est plus facilement visible quand cet effet est corrigé (données réduites), et l'indice d'assignement corrigé AI_c semble à cet égard beaucoup plus robuste que sa variance vAI_c et le F_{ST}.

Cet effet Wahlund nuit considérablement à l'image perçue au niveau de la structuration à l'échelle de la Suisse. Quand cet effet est contrôlé (au moins en grande partie), on observe un isolement par la distance très significatif, et les adultes non gorgés d'*I. ricinus* paraissent distribués en populations locales de tailles importantes (plus de 1 000 tiques par km²) et se dispersant difficilement à plus de 200 m par génération.

Il reste cependant bien d'autres questions et toutes ces hypothèses doivent être testées sur le terrain. Cette étude ouvre de nombreuses et prometteuses perspectives de recherche que je vous laisse le soin de discuter.

INTERACTIONS AVEC LES MICROPATHOGÈNES TRANSMIS

Introduction

La tique *I. ricinus* transmet un très grand nombre de pathogènes à ses multiples hôtes, dont la borréliose de Lyme qui, dans les régions boréales, représente un poids économique et en santé publique important (Gubler, 1998). Les agents de la borréliose de Lyme appartiennent au complexe d'espèces *Borrelia burgdorferi* sl. Il existe actuellement 12-13 espèces (ou génoespèces) reconnues dans le complexe dont 4, *B. burgdorferi*, *B. garinii*, *B. afzelii* et *B. spielmanii,* sont pathogènes pour l'homme (borréliose de Lyme) et deux autres, *B. valaisiana* et *B. lusitaniae*, sont de pathogénicité incertaine (Richter *et al.*, 2006 ; Postic *et al.*, 2007). Ces différentes espèces ne sont d'ailleurs pas responsables de symptômes identiques et présentent des spécificités d'hôtes réservoirs différentes (De Meeûs *et al.*, 2004b). En Europe de l'Ouest, *B. burgdorferi* est préférentiellement retrouvée chez l'écureuil roux, *B. afzelii* chez des campagnoles, des mulots et aussi l'écureuil roux, *B. garinii* plutôt chez des oiseaux et *B. spielmanii* uniquement chez le loir (Richter *et al.*, 2006). Quant à *B. valaisiana*, elle se montre spécifique des oiseaux, alors que *B. lusitanaiae* n'a été trouvée, en dehors de la tique *I. ricinus*, que chez un lézard (Postic *et al.*, 2007). L'épidémiologie de ces pathogènes reste largement mal connue et les résultats obtenus précédemment par nos analyses suscitent un certain nombre de questions. S'il y a spécificité différente des tiques immatures, sachant que les borrélies sont spécifiques des hôtes, les tiques des deux sexes devraient présenter des prévalences différentes pour les différentes espèces de

borrélies. En particulier, les femelles devraient porter davantage de borrélies d'hôtes peu mobiles (*B. burgdorferi*, *B. afzelii*) et les mâles celles d'hôtes plus mobiles (*B. garinii*, *B. valaisiana*). Ensuite, il est possible que l'infection par les borrélies puisse modifier le schéma de migration. Enfin, dans la mesure où un conflit/coopération pourrait exister au sein des tiques, existe-t-il une corrélation entre la présence des différentes espèces de borrélies au sein de tiques ?

Présentation des données

Toutes les tiques échantillonnées en Suisse pour cette étude avaient été coupées en deux, et une moitié envoyée à l'Institut de zoologie de Neuchâtel pour détermination de présence de borrélies et détermination de l'espèce (sondes moléculaires). L'autre moitié a été gardée dans l'alcool et un grand nombre utilisé pour génotypage microsatellite. Les données sont contenues dans le fichier TotBrutBorIR.txt où toutes les informations nécessaires sont disponibles. La présence ou l'absence de chaque espèce de borrélie trouvée est notée par un 1 ou un 0 dans la colonne correspondante. Un grand nombre de borrélies n'ont pu être déterminées au niveau de l'espèce (colonne "Bbundet") et seules trois espèces ont été trouvées : *B. burgdorferi* (Bbss), *B. afzelii* (Bba) et *B. garinii* (Bbg, trouvée trois fois).

Distribution des différentes borrélies dans les femelles et mâles d'*I. ricinus*

Pour cette analyse, nous allons devoir effectuer une régression logistique pour chaque espèce de borrélie (Bbundet, Bbss, et Bba). Bbg, trop rare sera laissée de côté. On va chercher à expliquer la présence de telle ou telle autre espèce de borrélie par le site, l'année et le sexe de la tique, ainsi que les interactions. Nous allons donc avoir besoin de R une fois de plus. Comme c'est le sexe que l'on souhaite tester ici, nous allons mettre ce facteur en premier (l'ordre compte dans les modèles de R). Après avoir lancé R et s'être positionné dans le répertoire approprié, on tape les commandes suivantes :

```
> data<-read.table("TotBrutBorIR.txt", header=TRUE)
> attach(data)
```

afin de faire lire l'ensemble du jeu de données à R (NB le > est automatiquement inséré par R). On spécifie ensuite le modèle en tapant la commande (sur une ligne) :

```
> Bba<-glm(data, formula =Bba ~ Sex + Site + Year + Sex:Site + Sex:Year +
Sex:Site:Year, family = binomial(link = logit))
```

On remarque que l'interaction entre facteurs est codée avec un ":" et que la régression est logistique, car on spécifie bien qu'elle appartient à la famille binomiale avec

un lien “logit” de la moyenne. Le lien logit signifie juste que la fonction qui relie la probabilité moyenne de la variable à expliquer (P_{Bba} probabilité de trouver une Bba) est du type log(P_{Bba}/(1 - P_{Bba})) et la variance égale à P_{Bba}/(1 - P_{Bba}). Dans notre cas, la variance est en fait inférieure à cette valeur et il y a sous-dispersion, ce dont nous discuterons plus loin.

Ensuite, il s’agit de tester le modèle par la commande :

```
> anova(Bba, test="Chi")
```

Le test est en effet un Chi2, car nous comparons des fréquences. Cette commande renvoie au résultat suivant :

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: Bba
Terms added sequentially (□rst to last)
                Df Deviance  Resid. Df  Resid. Dev             P(>|Chi|)
NULL                         857        358.68
Sex             1  0.32      856        358.36                 0.57
Site            7  35.69     849        322.66      8.290e-06
Year            1  8.84      848        313.83      2.951e-03
Sex:Site        7  10.32     841        303.51                 0.17
Sex:Year        1  0.82      840        302.69                 0.36
Sex:Site:Year   4  2.88      836        299.81                 0.58
Warning message:
In method(x = x[, varseq <= i, drop = FALSE], y = object$y, weights =
object$prior.weights,  :
des probabilités ont été ajustées numériquement à 0 ou 1
```

Nous constatons que seuls les termes “Site” et “Year” semblent importer et que le logiciel n’est apparemment pas très satisfait de la qualité des données. Pour simplifier ce modèle, une commande pratique est la commande “step” qui permet d’analyser la qualité de différents modèles plus simples en retirant et ajoutant des termes l’un après l’autre en commençant par les interactions d’ordre supérieur (celles faisant appel au plus grand nombre de facteurs). Ceci est évalué à l’aide d’un critère appelé AIC (*Akaike Information Criterion*) (Akaike, 1974) dont la valeur, qui doit être minimisée, est une mesure de la qualité d’ajustement du modèle statistique estimé par rapport aux données. Il ne s’agit pas d’un test, mais d’un outil d’aide à la sélection du modèle le plus simple permettant d’expliquer au mieux les données, le modèle doté du plus petit AIC étant le meilleur (cf. réponse 12 pour plus de précisions). En tapant donc la commande :

```
> step(Bba)
```

nous obtenons les résultats pour une série de différents modèles de plus en plus simples où les différents termes sont retirés un à un en commençant par l'interaction la plus complexe (Sex:Site:Year), qui est éliminée, l'AIC obtenu (338,69) s'avérant inférieur à celui du modèle complet (343,81), puis les interactions plus simples (Sex:Site et Sex:Year), jusqu'à ce que le retrait des facteurs conduisent à une augmentation de l'AIC par rapport au précédent. Ci-dessous sont présentés le début et la fin du processus :

```
Start:  AIC=343.81
Bba ~ Sex + Site + Year + Sex:Site + Sex:Year + Sex:Site:Year
                Df Deviance AIC
- Sex:Site:Year 4  302.69   338.69
<none>             299.81   343.81
Step:  AIC=338.69
Bba ~ Sex + Site + Year + Sex:Site + Sex:Year
                Df Deviance AIC
- Sex:Site      7  312.31   334.31
- Sex:Year      1  303.51   337.51
<none>             302.69   338.69
```

etc.

```
Step:  AIC=332.1
Bba ~ Site + Year
                Df Deviance AIC
<none>             314.10   332.10
- Year          1  322.96   338.96
- Site          7  345.43   349.43
Call: glm(formula = Bba ~ Site + Year, family = binomial(link = logit),
data = data)
```

La dernière ligne présentée ci-dessus donne le meilleur modèle. Suivent des informations sur les coefficients associés aux différents facteurs que nous n'allons pas utiliser, ainsi que des messages d'alertes sur la mauvaise qualité des données (on ne fait pas de miracles). Il s'agit maintenant d'analyser en détail ce meilleur modèle avec la série d'instructions (pour gagner du temps on peut copier le modèle ci-dessus et le coller après avoir tapé `"Bba2<-"`) :

```
> Bba2<-glm(formula = Bba ~ Site + Year, family = binomial(link = logit),
data = data)
> anova(Bba2, test="Chi")
```

qui renvoie au résultat :

```
Analysis of Deviance Table
Model: binomial, link: logit
```

```
Response: Bba
Terms added sequentially (first to last)
        Df Deviance  Resid. Df  Resid. Dev P(>|Chi|)
NULL                        857      358.68
Site     7  35.72           850      322.96 8.197e-06
Year     1  8.86            849      314.10 2.920e-03
```

La conclusion est donc qu'en ce qui concerne Bba, seuls le site et l'année importent. Ils expliquent respectivement 100×35,72/358,68 = 10 % et 100×8,86/358,68 = 2 % de la déviance totale. En procédant d'une manière identique pour Bbg, nous observons qu'aucune des variables n'explique les données alors que pour Bbss, en plus du site qui explique 28 % de la déviance totale (*P*-value < 0,001), le sexe des tique explique 3 % de la déviance (*P*-value = 0,007). Enfin, pour Bbundet le site seul explique 15 % de la déviance totale (*P*-value < 0,001).

Comme je l'ai déjà signalé plus haut, la dispersion des résidus ne suit probablement pas une loi binomiale et la variance est probablement différente de $P/(1 - P)$. Pour vérifier cela, il faut calculer le paramètre $\varphi = \text{Var}(\mu)\times(1-\mu)/\mu$ qui est ici inférieur à 1 (sous-dispersion) en particulier pour Bbss. On peut le calculer facilement avec la fonction "quasibinomial" (voir réponse 13). Comme seul Bbss a donné quelque chose de significatif pour le sexe des tiques, nous allons vérifier cela sur cette bactérie. Sous R, après avoir chargé le fichier de données si ce n'est déjà fait, nous allons taper les instructions suivantes :

```
> Bbss<-glm(data, formula =Bbss ~ Sex + Site, family =quasibinomial(link
= "logit"))
> summary(Bbss)
```

ce qui renvoie au résultat suivant (je ne garde que ce qui est le plus utile) :

```
Coefficients:
                        Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)            -20.31649    1194.11613     -0.017      0.9864
SexM                    -0.76071       0.31416     -2.421    0.0157 *
SiteCeneri               0.07671    2020.60021   3.80e-05      1.0000
SiteDorenaz             19.46080    1194.11614      0.016      0.9870
SiteEclepens            19.00830    1194.11616      0.016      0.9873
SiteGorges-du-Trient    16.48119    1194.11620      0.014      0.9890
SiteMontmollin          17.47997    1194.11624      0.015      0.9883
SiteNeuchâtel           17.08337    1194.11618      0.014      0.9886
SiteStaadswald           0.10793    1486.92130   7.26e-05      0.9999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasibinomial family taken to be 0.5155861)
```

Nous voyons donc que le le coefficient de dispersion est petit (0,52), il y a donc bien sous-dispersion (pour Bba ϕ = 0,93, il n'y a pratiquement pas de sous-dispersion

pour cette espèce-ci). Nous voyons également que le sexe des tiques est important (significatif) avec une estimation négative pour les mâles (les données partielles, corrigées des autres effets, sont centrées sur 0). Ceci est vérifiable en tapant la commande `anova(Bbss, test="F")` (les modèles quasi se testent avec un F), ce qui donne :

```
       Df  Deviance   Resid. Df   Resid. Dev   F        Pr(>F)
NULL                  857         286.155
Sex    1   7.233      856         278.922      14.029   0.0001922 ***
Site   7   80.730     849         198.192      22.369   < 2.2e-16 ***
```

Sachant que le comportement des modèles quasi en régression logistique peut s'avérer étrange quand l'événement étudié (présence de Bbss) est rare, ce qui est notre cas, on est en droit de chercher à renforcer ce résultat. En fin de compte, nous cherchons juste à vérifier si nous n'avons pas plus de Bbss chez les tiques femelles que chez les mâles, puisque ces borrélies sont spécifiques de petits rongeurs peu dispersants, supposés être davantage parasités par les larves et nymphes femelles que mâles, quel que soit le site ou l'année. On peut donc calculer parmi les tiques infectées par Bbss, la proportion de tiques femelles et mâles et comparer cette proportion à ½ par un test binomial. Sur 34 tiques infestées par Bbss, 26 étaient femelles, ce qui conduit à la *P*-value du test binomial (sous R, `binom.test(26, 34, p=0.5, alternative="greater")`) $P_{bino_26/34,0.5}$ = 0,0015, ce qui est équivalent aux résultats précédents. Vous vous demandez alors pourquoi vous ai-je cassé les pieds avec toutes ces régressions, alors qu'il était si simple de commencer par le test binomial ? La réponse est simple. D'abord, il n'est pas inutile d'apprendre à taquiner les régressions linéaires généralisées qui servent très souvent et, ensuite, dans une publication, une régression logistique en « quasi-likelihood » va avoir beaucoup plus de classe (en apparence) qu'un petit test binomial et impressionner beaucoup plus facilement ces referees désobligeants qui empoisonnent si souvent nos soumissions d'articles.

Donc Bbss, borrélie d'écureuil en Suisse, est plus fréquente chez les tiques adultes femelles que mâles, suggérant ainsi une prédisposition de ces femelles à se nourrir sur cet hôte quand elles sont aux stades larvaire et/ou nymphal.

Co-occurrence des différentes espèces de borrélies

Les différentes espèces de borrélies peuvent se retrouver en compétition, car elles partagent la même espèce de vecteur. Elles sont donc potentiellement en conflit et on pourrait s'attendre à un évitement. Au contraire, il pourrait y avoir association positive si les intérêts convergent ou si l'une des deux espèces immunodéprime ses hôtes et favorise ainsi l'entrée d'autres pathogènes. Il est donc intéressant de tester si ces borrélies se rencontrent au hasard ou non. La problématique est identique à une recherche d'association statistique entre deux états (infecté/non infecté) de

deux caractères (espèce x, espèce y). On peut donc simplement appliquer la même procédure que pour un test de déséquilibre de liaison. Il suffit donc de coder la présence de chaque borrélie comme un locus et l'absence par 11 et la présence par 22. Il y a donc quatre loci (Bba; Bbg, Bbss, Bbundet) avec chacun deux allèles (1 ou 2), toujours homozygotes (ou haploïdes). Pour ce faire, il suffit d'ouvrir le fichier "TotBrutBorIR.txt" et d'y remplacer, dans l'ordre, tous les 1 en 22 et tous les 0 en 11 et de fusionner les colonnes Site year sex pour obtenir quelque chose de la forme (fig. 43).

	Bbss	Bba	Bbg	Bbundet
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11
Bern95F	11	11	11	11

Figure 43
Début du fichier « TotBrutBorIRCoOccur.txt ».

Enregistrons ce fichier en le nommant "TotBrutBorIRCoOccur.txt" et importons-le dans Genetix afin de le convertir au format Fstat. Cliquez sur Fichier, Importer. Choisissez l'option fichier texte et double-cliquez sur "TotBrutBorIRCoOccur.txt". Choisissez les options séparateur tabulation, un chiffre par allèle, décochez la case de l'identifiant des individus et cliquez sur OK.

Il faut ensuite cliquer sur le menu Link. Dis et choisir Black & Kafsur comme sur la figure 44, ce qui aura pour effet de lancer une fenêtre de choix que vous devrez rendre comme dans la figure 45.

Cliquez ensuite sur OK et les résultats s'affichent dans TotBrutBorIRCoOccur.lkd.

Cliquez ensuite sur Outils, Conversion et FSTAT et nommez le fichier "TotBrutBorIRCoOccur.dat". Genetix construira donc un fichier où seront considérées comme appartenant à des populations différentes les tiques de sites, d'années et de sexes différents. Ouvrons ce fichier sous Fstat et sélectionnons les mêmes options qu'en figure 46.

Constatez que nous ne gardons que les fréquences alléliques (cela pourrait servir) et ce qui nous intéresse, le test de déséquilibre de liaison. On choisit dans un premier temps le niveau 5/100 pour aller plus vite. Cliquez sur "Run" et ensuite ouvrez le

D:\thierry\BouquinGenetPop\TP\DataIxodes\BorreliaGenetTicks\TotBrutBorIRCoOccur.txt

Fichier Edit AFC Link. Dis. Fstats Variabilité Distances Mantel Permutations Outils Aide

Black & Krafsur
Test sur permutations

		locus1	locus2	locus3	locus4
		Bbss	Bba	Bbg	Bbundet
Pop Ind1		001001	001001	001001	001001
Ind2		001001	001001	001001	001001
Ind3		001001	001001	001001	001001
Ind4		001001	001001	001001	001001
Ind5		001001	001001	001001	001001
Ind6		001001	001001	001001	001001
Ind7		001001	001001	001001	001001
Ind8		001001	001001	001001	001001
Ind9		001001	001001	001001	001001
Ind10		001001	001001	001001	001001
Ind11		001001	001001	001001	001001

Figure 44
Menu Link. Dis.

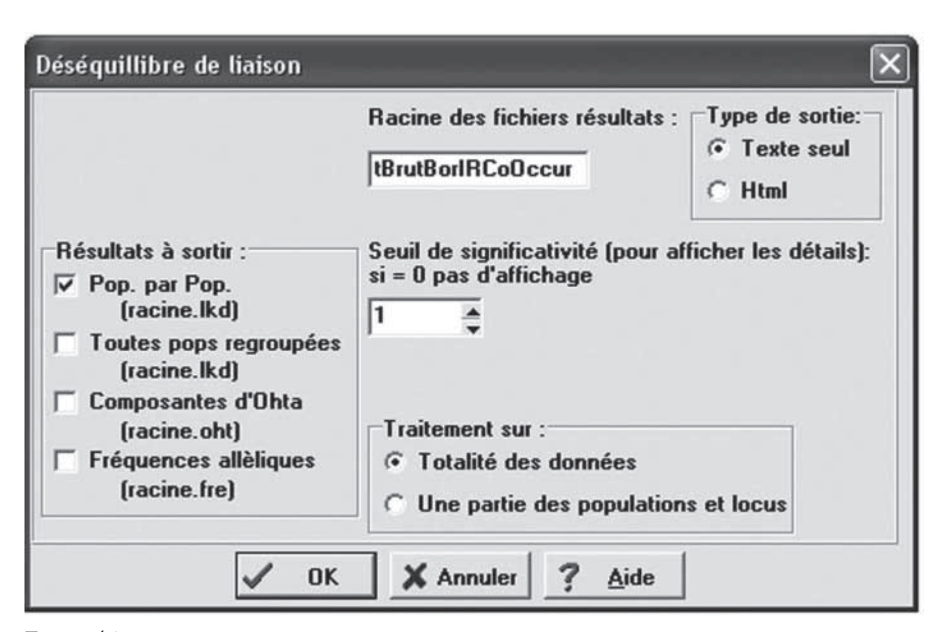


Figure 45
Choix à faire dans le menu de Black et Krafsur.

fichier "TotBrutBorIRCoOccur.out". Vous constatez que seulement 2 640 permutations ont été effectuées. Recommencez donc avec le niveau 1/100 pour le menu "Nominal level for multiple tests". Le résultat peut être synthétisé dans le tableau 18. On y voit clairement une association positive entre Bbss, Bba et Bbg, même si les

Figure 46
Menu Fstat pour tester l'association entre borrélies.

associations avec Bbg sont marginalement significatives, on peut considérer que le signal existe eu égard à la grande rareté de Bbg (puissance très faible du test). Il est intéressant de noter pour information que Bbundet, vraisemblablement composée d'une mixture de Bbg (très largement sous-représentée ici) et Bbv (*B. valaisiana* curieusement absente de l'échantillon) donnent des valeurs essentiellement négatives pour R(IJ), ce qui fait regretter plus encore que les déterminations de l'époque aient connu autant de problèmes. Il n'en reste pas moins qu'une forte corrélation positive lie Bbss, Bbg et Bba, qui est confirmée si on teste la co-occurrence des trois espèces dans la même tique rencontrée une fois dans l'échantillon des 73 tiques femelles de Neuchâtel en 1996, et pas à Bern comme annoncé dans la partie résultat de l'article de De Meeûs *et al.* (2004b) (on ne relit jamais assez ses épreuves). Il y a $N = 73$ observations, une fréquence observée de 4/73, 6/73 et 1/73 pour Bbss, Bbg et Bba respectivement, donc une fréquence attendue de $p = (4 \times 6 \times 1)/(73)^3$ pour l'événement de co-occurrence des trois borrélies dans la même tique, événement observé avec la fréquence $k = 1$. Cette fréquence observée peut être comparée à l'attendue par un test binomial. Sous R, tapez "`binom.test(1, 73, p=0.00006169, alternative="two.sided")`", ce qui donne une *P*-value = 0,0045 très significative. Cette *P*-value est en fait égale à la probabilité de l'événement lui-même puisqu'il n'y en a pas de plus rare possible. Elle est donc égale à la probabilité (dans une loi

binomiale) de tirer une seule fois Bbss-Bbg-Bba dans 73 tirages et où la probabilité de tirer Bbss-Bbg-Bba une fois (un tirage aléatoire) est de 0,00006169, soit (cf. n'importe quel manuel de statistiques) :

$$P_{Binomiale} = \binom{N}{k} p^k (1-p)^{(N-k)} = \frac{N!}{k!(N-k)!} p^k (1-p)^{(N-k)}, \text{ soit}$$

$$P_{Binomiale} = 73 \times 0{,}00006169 \times (1 - 0{,}00006169)^{72} = 0{,}0045$$

Tableau 18
Valeurs (moyennes pondérées par les tailles de sous-échantillons) de corrélation entre la présence de chacune des deux bactéries considérées (ligne 2-2 dans la sortie de Genetix, colonne R(IJ)) et *P*-value (test *G* multi-échantillon de Fstat) correspondante. Une valeur de *R*(IJ) positive indique une association positive, alors qu'une valeur négative indique une répulsion.

Borrélies (I x J)	*R*(IJ)	*P*-value
Bbss × Bba	0,292	0,00008
Bbss × Bbg	0,496	0,05311
Bbss × Bbundet	- 0,069	1
Bba × Bbg	0,109	0,09348
Bba × Bbundet	- 0,017	0,91598
Bbg × Bbundet	- 0,030	1

Cette corrélation est donc très forte. Elle peut être due au fait que les tiques infectées correspondent à des individus sensibles et que les autres individus sont résistants. Cette corrélation peut également provenir du fait qu'être infecté par une des trois borrélies tend à favoriser l'infection par les deux autres (par immunosuppression, par exemple). Ceci peut être testé en ne regardant que les tiques infectées. La corrélation existe-t-elle toujours ?

Nous allons donc recommencer la même opération, mais avec un fichier ne contenant que les tiques infectées par au moins une borrélie (donc sensibles). Le résultat de cette analyse est présenté dans le tableau 19. L'image qui en ressort est totalement différente puisque les associations sont presque toutes négatives, témoignant d'une exclusion entre espèces de borrélies au sein des tiques avec deux exclusions très significatives entre Bbundet (vraisemblablement composée de Bbg et Bbv) et les deux borrélies les plus fréquentes Bbss et Bba (*R*(IJ) ≈ - 0,6, *P*-value ≤ 0,003). Il y a donc dans les populations d'*I. ricinus* des tiques plus sensibles ou plus exposées que d'autres aux borrélies de toutes espèces où ces dernières s'accumulent. Par contre, au sein de ces tiques plus sensibles, ou davantage exposées, les différentes espèces de borrélies tendent à « s'éviter », en particulier

Bbundet, que l'on peut raisonnablement soupçonner d'être essentiellement composée de borrélies d'oiseaux (Bbg et Bbv) et les borrélies de micromammifères (Bbss et Bba). Ceci relance le débat sur l'existence de spécificités différentielles des tiques et donc de races d'hôtes, en particulier vis-à-vis des oiseaux et des micromammifères, au moins aux stades larve et nymphe. Il pourrait s'agir aussi d'une manipulation des tiques par les borrélies qui « inciteraient » une tique infectée par telle ou telle autre borrélie à faire son prochain repas sur la même espèce hôte que celle l'ayant infectée. Nous aurons l'occasion de discuter davantage cela dans le paragraphe suivant.

Tableau 19
Valeurs (moyennes pondérées par les tailles de sous-échantillons) de corrélation entre la présence de chacune des deux bactéries considérées (ligne 2-2 dans la sortie de Genetix, colonne *R*(IJ)) et *P*-value (test *G* multi-échantillon de Fstat) correspondante, mais avec les tiques infectées par au moins une borrélie uniquement (tiques saines exclues du test). Une valeur de *R*(IJ) positive indique une association positive, alors qu'une valeur négative indique une répulsion.

Borrélies	*R*(IJ)	*P*-value
Bbss × Bba	- 0,471	0,11049
Bbss × Bbg	0,412	1
Bbss × Bbundet	- 0,622	0,00304
Bba × Bbg	- 0,212	0,67559
Bba × Bbundet	- 0,632	0,0001
Bbg × Bbundet	- 0,290	1

Occurrence des différentes espèces de borrélies et génétique des tiques

Dans cette partie, nous rechercherons s'il existe une relation entre la génétique des tiques et leur probabilité d'infection par chacun des quatre types de borrélies. On peut répondre à cette question de trois manières. Soit en testant la différenciation génétique entre tiques infectées et non infectées dans chaque sous-échantillon, soit en testant la différenciation, dans chaque sous-échantillon, entre tiques infectées par des borrélies différentes, enfin en procédant à un test de biais de structuration, comme nous l'avons fait pour le sexe des tiques, mais avec le statut infecté/non infecté à la place.

Différenciation entre tiques infectées et non infectées

Il faut construire un fichier par espèce de bactérie Bbss, Bba et Bbundet (il n'y a pas assez de Bbg). On doit changer de nom de population pour chaque site, année et sexe. La figure 47 donne un exemple de fichier pour Bbss.

	IR08	IR25	IR27	IR32	IR39
Dor-96-F-I	170184	000000	119119	233246	129142
Dor-96-F-I	165168	147150	119119	235250	127142
Dor-96-F-I	168168	145146	119119	233250	124127
Dor-96-F-I	169175	146146	123123	233246	128128
Dor-96-F-I	168176	149149	119119	250250	127138
Dor-96-F-I	172175	148148	121123	235235	142142
Dor-96-F-I	174175	146146	119119	233233	135135
Dor-96-F-I	168170	147153	119119	241241	131131
Dor-96-F-I	167168	141143	119119	233248	139139
Dor-96-F-I	173177	000000	123123	233241	128142
Dor-96-F-I	176176	144144	119119	246246	135135
Dor-96-F-I	166172	145149	121121	233246	138138
Dor-96-F-I	175183	130147	119119	246250	000000
Dor-96-F-I	172174	134134	119119	233250	000000
Dor-96-F-U	173177	141154	119119	235245	125142
Dor-96-F-U	170175	143148	119123	233235	000000
Dor-96-F-U	168174	145147	119119	241241	142142
Dor-96-F-U	167176	135135	117117	250250	142142
Dor-96-F-U	168174	149150	119119	000000	127128
Dor-96-F-U	179182	130130	119119	233233	129135
Dor-96-F-U	173175	149149	119119	233243	135142
Dor-96-F-U	167183	144144	123123	233250	127128
Dor-96-F-U	181181	146146	125125	246250	128128

Figure 47
Exemple des premiers individus du fichier de données pour tester la différentiation entre tiques infectées (I) et non infectées (U) par Bbss. Nous voyons ici les tiques de Dorénaz 1996 femelles.

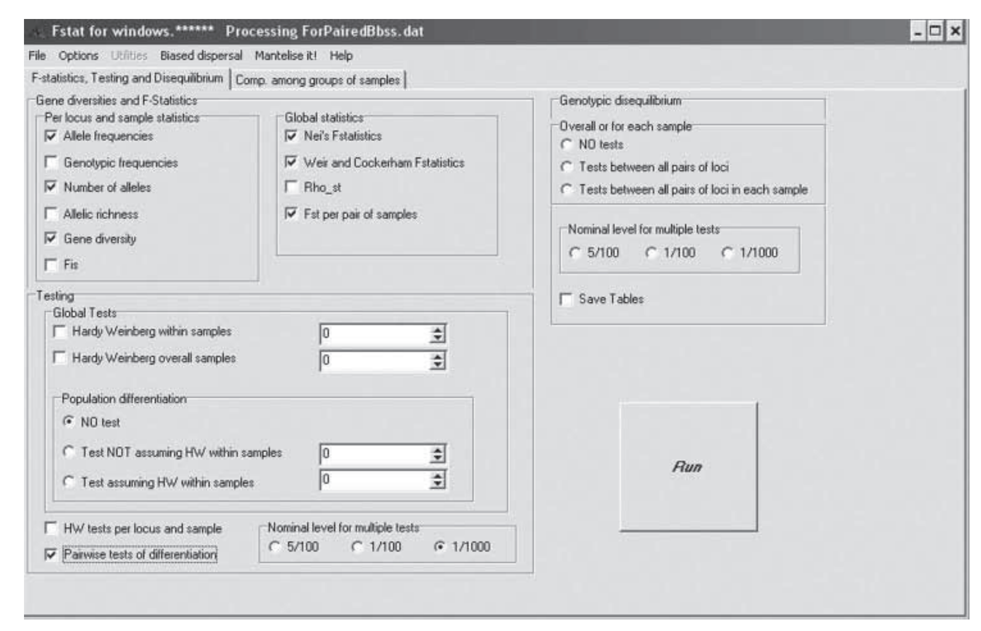


Figure 48
Exemple des cases à cocher pour une analyse de différenciation par paire d'échantillons, exemple des tiques infectées ou non par Bbss.

Il suffit ensuite de convertir ce fichier au format Fstat (en passant par Genetix, par exemple) et de procéder sous Fstat au calcul des F_{ST} par paire de sous-échantillons et au test de différenciation par paire, comme indiqué dans la figure 48. Vous constatez que j'ai coché la case 1/1000 pour le nominal level afin d'obtenir au moins 10 000 permutations et donc d'obtenir des *P*-values assez précises. Le fichier de données s'appelle "ForPairedBbss.dat" et les fichiers de sortie qui nous intéressent sont "ForPairedBbss.fst" pour récupérer les valeurs de F_{ST} par paire qui nous intéressent et "ForPairedBbss-pp.pvl" où nous allons récupérer les *P*-values correspondantes. Attention, dans ces fichiers, seules les comparaisons entre tiques infectées et non infectées du même sexe, de la même année et du même site nous intéressent. Le résultat pour Bbss est présenté dans le tableau 20.

Tableau 20
Compilation des résultats obtenus lors de l'analyse de la différenciation entre paires de sous-échantillons infectés et non infecté par Bbss. La combinatoire est obtenue par la moyenne non pondérée des F_{ST} et un test binomial généralisé pour les *P*-values.

Sous-échantillon	F_{ST}	*P*-value
Dorénaz 1996 femelles	- 0,008	0,6477
Dorénaz 1996 mâles	- 0,030	0,3226
Eclepens 1996 femelles	0,008	0,1206
Eclepens 1996 mâles	0,027	NA
Gorges-du-Trient 1996 femelles	- 0,034	0,9171
Montmollin 1996 mâles	- 0,027	NA
Neuchâtel 1996 femelles	- 0,001	0,7250
Combinatoire	- 0,009	0,5179

Vous remarquerez que la combinaison des cinq tests disponibles a été effectuée à l'aide de la procédure binomiale généralisée de Teriokhin *et al.* (2007) effectuée à l'aide du logiciel MultiTest (De Meeûs *et al.*, 2009). En effet, à partir de quatre tests, je préfère utiliser cette procédure plutôt que le test *Z* de Stouffer (Whitlock, 2005). Pour effectuer ce dernier, chaque *P*-value individuelle est transformée en son équivalent de la distribution *Z* centrée sur 0 et d'écart-type 1. Sous Excel, on tape `=SI(B2="NA";"";SI(B2>0.9999;LOI.NORMALE.INVERSE(0.9999;0;1);LOI.NORMALE.INVERSE(B2;0;1)))`. B2 correspond aux coordonnées de la case du tableau Excel où la *P*-value à transformer se trouve. Cette commande renvoie une absence de résultat quand "NA" est rencontré et tient compte du fait qu'une *P*-value de 1 n'est pas transformable et la *P*-value = 0,9999 est choisie comme limite supérieure. Enfin,

l'équivalent de la *P*-value en Z centrée réduite de moyenne 0 et d'écart-type 1 est calculé. Les valeurs Z_i obtenues sont ensuite combinées dans la formule (Whitlock, 2005) :

$$Z_s = \frac{\sum_i^k Z_i}{\sqrt{k}}$$, où k est le nombre de tests (= 0,3266 ici).

La *P*-value globale s'obtient ensuite par un retour à la loi normale, soit sous Excel : `=LOI.NORMALE.STANDARD(Zs)`(=0,628 ici). Vous trouverez un argumentaire plus détaillé dans De Meeûs *et al.* (2009) pour les situations où la procédure binomiale généralisée ou le test Z doivent ou peuvent être utilisés.

Si on procède de la même façon pour Bba et Bbundet, le même type de résultat émerge, même quand on ne distingue pas le sexe des tiques (échantillons plus grands) puisque pour ces deux catégories de borrélies, nous avons vu que le sexe des tiques n'importait pas. Ce résultat est rassurant car, étant donné que les marqueurs sont non codants (donc neutres) et indépendants, il eut été difficile d'interpréter une différenciation entre tiques infectées et non infectées, à moins d'évoquer l'existence d'espèces cryptiques de tiques et une spécificité des borrélies.

Différenciation entre tiques infectées par différentes borrélies

Ici, il faut ne garder que les tiques infectées et définir comme sous-population les tiques du même sexe, échantillonnées la même année, dans le même site et ayant le même statut infectieux. Notons qu'une tique infectée par Bba et Bbss ne fera pas partie de la même sous-population qu'une tique infectée par Bbss seule. On met ensuite le fichier au format Fstat et on lance la procédure de F_{ST} par paire. Ce faisant, vous constaterez que la plupart des tests sont infaisables, c'est normal. Les résultats sont compilés dans le tableau 21. En toute rigueur les tests, qui ne sont pas tous indépendants, devraient subir la correction de Bonferroni. Cependant, étant donné la faiblesse des échantillons (manque total de puissance), nous nous abstiendrons de le faire. Le seul F_{ST} positif est obtenu entre Bba et Bbundet, mais il n'est pas significativement plus grand que 0. Eu égard à la faiblesse des tailles de sous-populations ici, nous décidons que rien ne permet d'affirmer l'existence d'une différence génétique entre tiques infectées par différentes bactéries et rien ne permet de l'exclure formellement au moins pour ce qui concerne le couple Bba/Bbundet. S'il existe des races d'hôtes chez *I. ricinus*, ce n'est pas avec ces données qu'on peut le montrer.

Biais de structuration spécifique associé au pathogène

Ici, il faut reprendre les données pour chaque espèce de bactérie et créer un fichier de type Genepop comme ce qui a été fait en p. 147-150, sauf qu'ici les tiques sont distinguées en fonction de leur statut infectieux et non par leur sexe, tel que dans la figure 49. Notons que nous ne traitons que les sites prélevés en Suisse et où au moins une tique infectée est trouvée. Parce qu'il y a un biais de structuration sexe-spécifique,

Tableau 21
Compilation des résultats des tests de différenciation, parmi les tiques infectées, par paire en fonction de l'espèce de bactérie présente et pour les paires effectivement trouvées. Quand plusieurs tests indépendants sont disponibles ils sont combinés : les F_{ST} sont des moyennes non pondérées, alors que les *P*-values ont été obtenues par la procédure *Z* (il y a en effet systématiquement moins de quatre tests ici).

Borrélies	Sous-échantillon	F_{ST}	*P*-value
Bbss/Bba	Dor96F	- 0,0095	0,8577
	Gor96F	0,0000	0,6628
	Combinés	- 0,0047	0,8540
Bba/Bbundet	Mon96F	- 0,0357	1
	Sta96F	0,1025	0,0662
	Sta96M	0,0454	0,1687
	Combinés	0,0374	0,7657
Bba/Bbss+Bba	Dor96F	- 0,0501	0,8560
Bbss/Bbss+Bba	Dor96F	0,0004	0,5998
Bba/Bba+Bbundet	Sta96M	0,0269	0,0676
Bbundet/Bba+Bbundet	Sta96M	- 0,0394	0,8043

+ signifie la co-occurrence de deux espèces de borrélies

ainsi que des différences d'infection, les femelles et les mâles sont analysés séparément. Cependant, parce que la taille des échantillons est très faible (peu de borrélies trouvées et identifiées), nous combinerons le tout dans un seul fichier (gain de puissance). On prendra soin de distinguer les tiques d'années et de sexe différents comme appartenant à des populations différentes (séparées par un “pop” dans le fichier).

Quand le fichier est constitué, il faut ensuite lancer Fstat et cliquer sur le menu “Biased dispersal”. La fiche correspondante apparaît alors. Il faut ensuite charger le fichier à analyser en cliquant le menu “File” et “Open” et cocher les cases comme en figure 50 puis sur le bouton “Go!”. Pour une raison que j'ignore, il faut cocher tous les paramètres si on souhaite obtenir le résultat du test sur H_s, en particulier F_{IS} et H_o qui ne sont guères utiles ici, car nous avons codé les mâles homozygotes pour IR08.

Le résultat est contenu dans un fichier de type nomdufichier.res (un fichier par espèce de borrélie). Le résultat principal concerne le test du F_{ST} (et aussi la relatedness, ce qui est normal si on regarde sa définition dans la documentation de Fstat) et est présenté dans le tableau 22.

```
3·5·99¶
IR08,·IR25,·IR27,·IR32,·IR39¶
Pop¶
I,Dorenaz,·7070·4247·2323·4650·3742¶
I,Dorenaz,·6868·3434·1919·4343·4144¶
U,Dorenaz,·7575·4853·1919·3350·3232¶
I,Dorenaz,·7474·4345·1919·3333·0000¶
U,Dorenaz,·6868·4848·1919·0000·3042¶
U,Dorenaz,·7878·3441·1919·3535·3030¶
U,Dorenaz,·7272·4244·1919·3535·2835¶
U,Dorenaz,·7777·3446·1919·5050·2627¶
U,Dorenaz,·7575·4848·1919·4650·2931¶
U,Dorenaz,·8383·4747·2525·3339·3636¶
U,Dorenaz,·6969·4242·1921·0000·2729¶
U,Dorenaz,·7575·4247·1919·3535·4242¶
U,Dorenaz,·7676·0000·1919·4650·2929¶
U,Dorenaz,·7373·0000·1919·3333·2528¶
U,Dorenaz,·0000·4545·0000·0000·2035¶
U,Dorenaz,·7373·3030·1919·5050·3535¶
U,Dorenaz,·0000·0000·0000·3333·0000¶
Pop¶
U,Eclepens,·0000·4747·1919·3343·3737¶
U,Eclepens,·7676·4848·0000·5050·0000¶
U,Eclepens,·7171·3648·1919·0000·3336¶
U,Eclepens,·7676·4646·1919·0000·2842¶
```

Figure 49
Type de données pour le test de biais de structuration pathogène spécifique.
Les tiques infectées sont notées avec un I et les saines avec un U.
Se référer à l'aide de Fstat pour plus de détails sur la constitution d'un tel fichier.
Il est important de ne pas oublier que les allèles doivent être à deux chiffres
et que les colonnes sont séparées par des espaces et non des tabulations.

Tableau 22
Résultat du test basé sur le F_{ST} de biais de structuration génétique pathogène spécifique des tiques pour les différentes espèces de borrélies pour lesquelles assez de données étaient disponibles (Bbg exclue). On remarque une structuration significativement plus forte pour les tiques infectées (I) par Bba par rapport aux tiques non infectées par cette borrélie (U).

	Bbss	Bba	Bbundet
U	0,001	0,002	0,000
I	- 0,015	0,076	- 0,045
P-value	0,4998	0,0033	0,1764

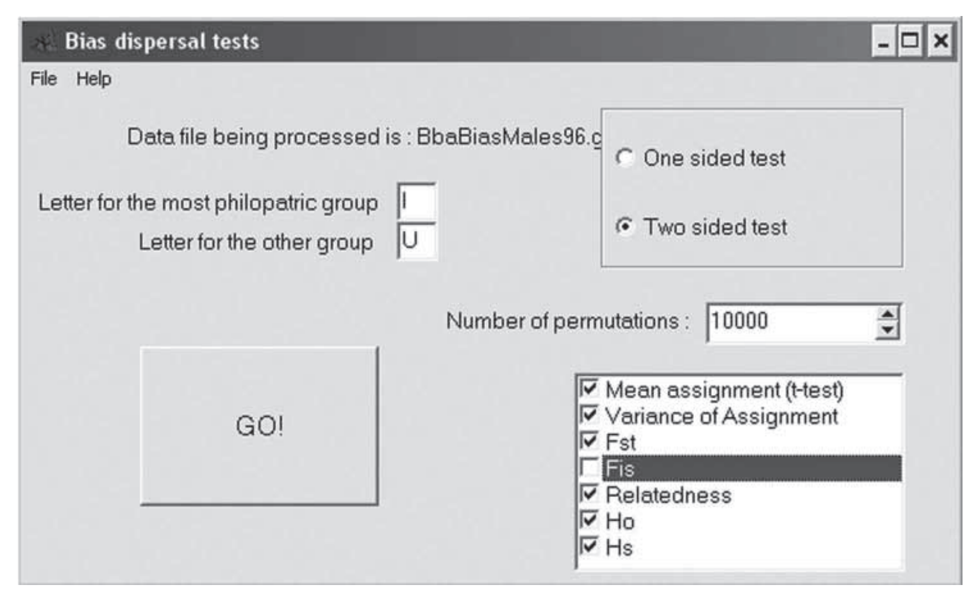


Figure 50
Cases à cocher pour l'analyse du biais de dispersion pathogène spécifique.
Le test demandé est bilatéral, car nous n'avons en principe pas d'à priori.
Toutes les cases sont cochées, même les cases "Fis" et "Ho"
(inutiles à cause du codage de IR08), car on souhaite obtenir le résultat pour H_s.

Il y a donc bien un biais de structuration dû à l'infection par Bba. Plusieurs hypothèses peuvent expliquer ce résultat. La première hypothèse implique que certaines tiques, plus sensibles à l'infection par Bba sont aussi pléiotropiquement moins mobiles. Les marqueurs utilisés étant des microsatellites non codants, cela impliquerait un déterminisme génomique peu vraisemblable. Par ailleurs, l'absence totale de différenciation entre tiques infectées et non infectées (montrée en p. 179-182) discrédite cette interprétation.

La deuxième hypothèse implique l'existence d'au moins deux espèces cryptiques dont l'une, moins mobile que la seconde, serait plus sensible à l'infection par Bba. Notons que nous n'avons noté aucun déséquilibre de liaison (attendu en pareil cas). Par ailleurs, si on calcule avec Fstat le F_{IS} des tiques en séparant celles infectées par Bba de celles qui ne le sont pas, on ne retrouve aucune diminution du F_{IS} (~ 0,45 pour les infectées et ~ 0,44 pour les non infectées par Bba). Il n'existe pas de différenciation significative entre tiques infectées et non infectées. Cette interprétation n'est donc aucunement soutenue.

La troisième interprétation possible impliquerait l'existence d'une adaptation locale des borrélies qui infecteraient plus facilement les tiques locales (résidentes) que les immigrantes. Deux arguments vont à l'encontre de cette hypothèse.

La première est que les tiques mâles et femelles qui en principe n'ont pas la même dispersion (les femelles dispersent en principe peu ou pas, cf. p. 147-153) ne sont pas infectées différemment par Bba (p. 173). Par ailleurs, c'est le partenaire le plus mobile des deux qui doit en théorie être le mieux adapté localement (Gandon *et al.*, 1996 ; Gandon, 2002). Or ici, les tiques sont modestement structurées alors que l'on pense que les borrélies le sont beaucoup plus (Qiu *et al.*, 1997). C'est donc l'hôte (la tique) qui devrait être adapté localement et non l'inverse.

La quatrième hypothèse implique une survie plus faible des tiques migrantes quand ces dernières sont infectées par Bba. Comme les tiques femelles sont moins mobiles que les mâles, ce sont ces derniers qui devraient être les plus affectés par ce phénomène. Ceci est testable en refaisant l'analyse sur les tiques femelles et mâles séparément. Cela suppose une survie au stress moins bonne des larves et/ou nymphes infectées par Bba.

La cinquième hypothèse est la plus séduisante. Elle implique une manipulation des larves et nymphes par la borrélie. Cette borrélie est spécifique de petits rongeurs. Il est donc plus intéressant pour elle d'être injectée dans un petit rongeur, peu dispersant, que dans un oiseau ou un grand mammifère, hôtes beaucoup plus mobiles. Les Bba capables de manipuler les tiques qu'elles infectent de sorte que ces dernières préfèrent se fixer sur un petit rongeur plutôt que sur d'autres hôtes seraient donc avantagées. Cette hypothèse est testable en laboratoire, mais cela n'a malheureusement jamais été fait. Cela implique aussi, comme pour l'hypothèse précédente, que les femelles, déjà très peu mobiles, seront moins affectées par le biais de structuration Bba-spécifique que les mâles.

Biais de structuration spécifique au pathogène et au sexe

Nous allons utiliser la même procédure que précédemment, mais en divisant le fichier en deux : un fichier pour les tiques femelles et un autre pour les tiques mâles. Cette fois, les tests seront faits de manière unilatérale avec I (infectés) comme catégorie la plus philopatrique. Il y a deux raisons à cela. La première est que l'on connaît d'avance le sens du signal. La seconde raison est que les échantillons étant encore plus petits, nous aurons besoin d'encore plus de puissance dans le test. Nous ne nous occuperons que du test sur le F_{ST}. Les tests sont tous les deux significatifs avec P-value = 0,0497 pour les tiques femelles et P-value = 0,0123 pour les tiques mâles et une apparente très forte différence de signal entre les deux, comme indiqué dans la figure 51.

Nous pouvons également constater la formidable variance chez les mâles infectés (très peu nombreux). Nous pouvons effectuer un test unilatéral de Wilcoxon pour données appariées comme en p. 143 quand nous avions comparé les F_{IS} des données brutes avec ceux des données clusterisées par BAPS. Ici, l'unité d'appariement reste le locus (donc cinq données), mais la statistique est la différence de F_{ST} entre tiques

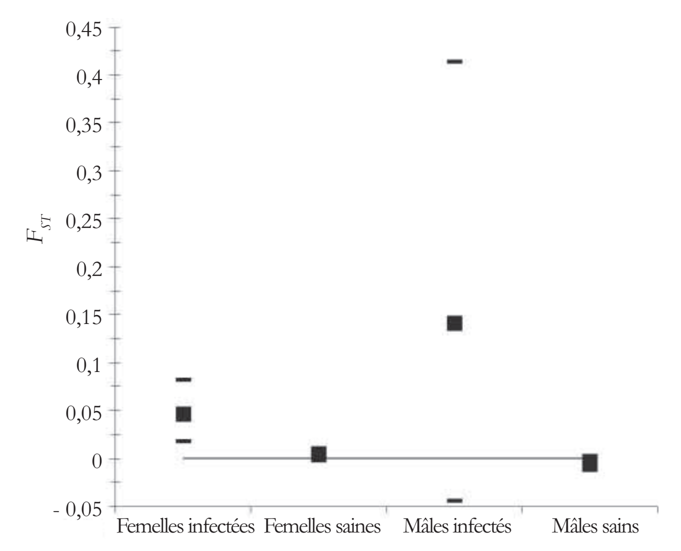


Figure 51
Différence comparée entre tiques mâles et femelles du F_{ST} mesuré entre tiques infectées par Bba et celles qui ne le sont pas.

infectées et non infectées chez les femelles et les mâles. Le fichier à tester contiendra donc les différences des différences appariées : (F_{STMI}-F_{STMU})-(F_{STFI}-F_{STFU}). Le test unilatéral ($F_{ST\mathrm{MI}}$-$F_{ST\mathrm{MU}}$>$F_{ST\mathrm{FI}}$-$F_{ST\mathrm{FU}}$) montre que la différence n'est pas significative, même si la *P*-value reste relativement faible (0,17). Ceci illustre les limites de notre jeu de données (beaucoup trop de données manquantes).

CONCLUSIONS SUR LES BORRÉLIES ET *I. RICINUS* EN SUISSE

Au cours de nos analyses, nous avons constaté que Bbss, borrélie d'écureuil, était plus souvent retrouvée chez les tiques mâles que femelles, ce qui est attendu si, comme le suggérait le biais de dispersion sexe-spécifique détecté chez ces tiques, les larves et nymphes femelles préfèrent se nourrir sur des rongeurs (peu dispersants). Rien de tel n'a pu être trouvé pour Bba pour laquelle ceci était attendu également, peut-être parce qu'une certaine quantité de tiques infectées par cette borrélie fait partie du stock Bbundet. Quant à Bbg, trop rarement détectée, d'autres études seront requises afin de déterminer si, comme attendu, elle est plus souvent retrouvée chez les tiques mâles.

Certaines tiques sont plus sensibles ou plus exposées à l'infection par les borrélies en général, comme l'attestent les fortes corrélations positives observées sur les co-occurrences des trois espèces Bbss, Bba et Bbg. En se concentrant sur ces tiques sensibles (infectées par au moins une borrélie), il y a un évitement manifeste. Les corrélations deviennent toutes négatives, exception faite de l'association Bbss×Bbg, pour qui le faible nombre de Bbg détectées rend les choses difficiles à interpréter, et très significatives pour les couples Bba×Bbundet et Bbss×Bbundet. Cette dernière observation peut laisser spéculer que ces borrélies indéterminées soient majoritairement des borrélies d'oiseaux (Bbg et Bbv) très déficitaires dans notre jeu de données. Dans ce cas, nous pourrions proposer que les larves et nymphes sensibles se subdivisent en tiques ne se nourrissant que sur une gamme limitée d'hôtes réservoirs de borrélies spécifiques. Tout dépend de l'identité spécifique de ces Bbundet. Les données ne permettent pas d'exclure l'existence d'une telle spécificité en races d'hôtes. La manipulation de la spécificité des tiques par les borrélies ne peut pas non plus être exclue. C'est aussi cette manipulation qui expliquerait le biais de structuration des tiques infectées par Bba. D'une manière générale, on ne peut que regretter le nombre de données manquantes qui limite nos conclusions mais aussi remarquer que, malgré cela, de nombreuses perspectives nouvelles de recherche ont émergé qui illustrent la puissance des outils offerts par la génétique des populations.

5 Glossina palpalis gambiensis le long de la rivière Mouhoun au Burkina Faso

INTRODUCTION

Ce jeu de données a fait l'objet d'un article (Bouyer *et al.*, 2009). Il permettra de réviser plusieurs notions et de mettre en pratique de nouvelles méthodes d'analyse telles que l'isolement par la distance entre individus. Comme pour les tiques, ce jeu de données est téléchargeable, mais dans un format différent. Ce fichier s'appelle "TsetseJerCoordGeo&Trap&SexTotData.xls". Dans ce chapitre, nous ne détaillerons que les analyses nouvelles et irons la plupart du temps très vite sur les notions déjà illustrées dans le précédent chapitre.

ÉTAT DES LIEUX

Les trypanosomoses africaines figurent parmi les plus sérieuses des maladies tropicales négligées (Schofield et Kabayo, 2008). L'OMS estime que le nombre total de cas de maladie du sommeil avoisine les 300 000 personnes (WHO, 2006a). Par ailleurs, d'après la FAO, le coût économique des trypanosomoses animales (nagana) atteint 4,74 milliards de dollards US par année (FAO, 2000). En 2001, plusieurs pays africains ont lancé le Pan African Tsetse and Trypanosomosis Eradication Campaign (PATTEC) afin d'établir une lutte concertée contre cette plaie à l'origine de nombreux problèmes de faim, d'appauvrissement et de frein au développement d'une agriculture durable dans les zones rurales d'Afrique subsaharienne (http://www.africa-union.org/Structure_of_the_Commission/depPattec.htm). *Glossina palpalis* s.l. (une des espèces de mouches tsé-tsé) est un des plus importants vecteurs de trypanosomoses humaine et animales en Afrique de l'Ouest. En Guinée, la sous-espèce *Glossina palpalis gambiensis* (Gpg) transmet la maladie du sommeil avec une prévalence relativement élevée (Camara *et al.*, 2005). Au Burkina Faso, c'est un vecteur majeur de nagana, en particulier dans le bassin de la rivière Mouhoun où se situent les échantillons que nous allons analyser (Bouyer *et al.*, 2006). La connaissance des schémas de dispersion et de tailles de populations est un pré-requis nécessaire au développement d'une lutte raisonnée pour le contrôle des populations de vecteurs (Tabachnick et Black, 1995). Pour les mouches tsé-tsé, comme pour les autres espèces de vecteurs, les estimations directes par marquage-recapture sont fastidieuses et coûteuses et pas nécessairement très fiables dans le cas des mouches tsé-tsé (Terblanche et Chown, 2007). Comme nous allons

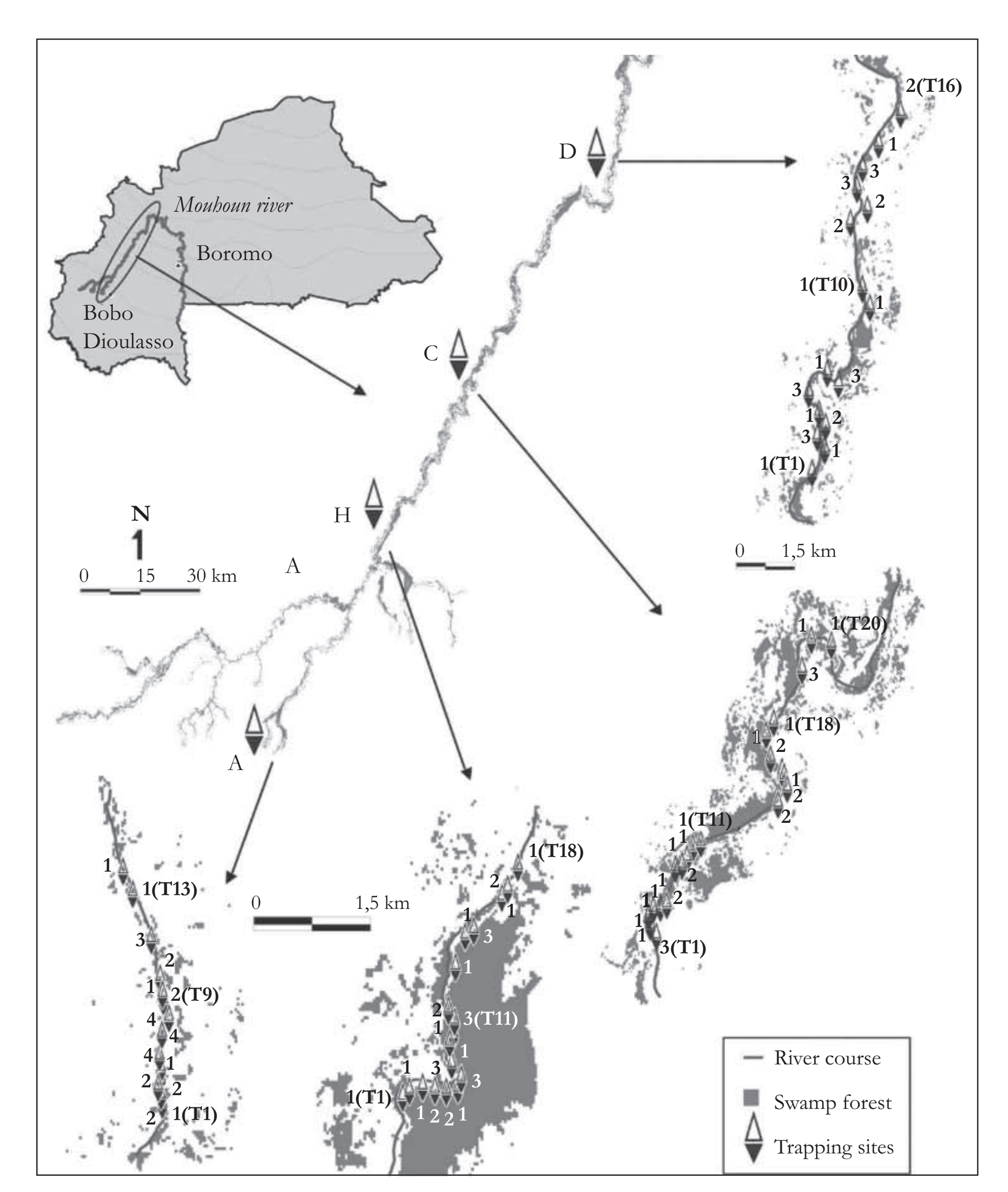


Figure 52
Les quatre zones d'échantillonnages (A, H, C et D) et la localisation précise de chaque piège dans chaque zone de captures de Gpg le long du Mouhoun. Pour chaque piège (bicone bleu), le nombre de glossines génotypées est donné. Le rang des pièges de chaque zone est donné entre parenthèses pour les premier, dernier pièges et intermédiaire (d'après Bouyer *et al.*, 2009).

le voir, les marqueurs génétiques et les outils de la génétique des populations peuvent apporter une solution très efficace.

Le jeu de données concerne des échantillons de Gpg prélevées dans quatre zones le long de la rivière Mouhoun (fig. 52).

PREMIER RECODAGE DES DONNÉES

Les données brutes se présentent comme dans le tableau 23. La première colonne indique le site de prélèvement (A, H, C ou D, comme dans la figure 52). Les deuxième et troisième colonnes correspondent aux coordonnées GPS des pièges suivies du nom du piège en quatrième colonne, du sexe et du nom des individus glossines génotypés en colonnes cinq et six respectivement. Suivent les génotypes des allèles aux sept loci étudiés avec une colonne par allèle et donc 14 colonnes (colonnes 7 à 21). Vous remarquerez que les loci liés à l'X possèdent cette lettre dans leur nom (comme pour PgpX11, par exemple) et que les mâles ont été codés homozygotes pour ces loci, ce dont il faudra se souvenir au moment de tester la panmixie. Les données manquantes sont, quant à elles, codées par des "0". Pour tous les tests liés à l'hétérozygotie locale, il faut créer un second fichier "TsetseJerCoordGeo&Trap&SexTotDataMalManq.xls" où les mâles sont manquants aux loci liés à l'X.

Ensuite, nous allons utiliser un nouveau logiciel très pratique qui peut convertir facilement nos deux fichiers dans des formats variés, y compris pour les programmes dont nous avons besoin. Ce programme s'appelle Create v 1.1 (Coombs *et al.*, 2008). Vous lancez Create et remplissez la fiche comme dans la figure 53.

Quand vous sélectionnez le fichier Excel, le programme vous demande dans quelle fiche Excel[12] se trouvent les données. Cliquez sur celle qui convient (la 1 en principe). Cliquez ensuite sur "Proceed". Le programme vous demande de vérifier qu'il a bien pris en compte ce qu'il fallait en vous montrant l'exemple du premier individu. Répondez oui si ça colle. Un second menu apparaît qu'il vous faut remplir comme en figure 54. Vous obtenez ainsi quatre fichiers, deux pour les données en format Genepop et Fstat, et deux pour le nom des populations. Faites la même chose pour "TsetseJerCoordGeo&Trap&SexTotDataMalManq.xls". Nous allons dans un premier temps tester les déséquilibres de liaison avec "TsetseMouhouMalHomo-FSTAT.dat" et les F_{IS} avec "DataTsetseMouhounMalManq-FSTAT.dat", fichiers Fstat que vient de créer Create. Vous pouvez renommer ces fichiers avec des noms moins longs. Vous pouvez aussi éditer les fichiers *.lab et supprimer les colonnes supplémentaires qu'a créé Create (je ne sais pas pourquoi il fait ça)

[12] On peut aussi charger un fichier de données au format texte seul.

Tableau 23
Extrait du jeu de données brutes des génotypes des individus Gpg capturées le long de la Mouhoun. Le tableau est tronqué pour les derniers loci. Notez qu'un locus occupe deux colonnes.

Site	Longitude	Latitude	Piège	Sexe	Individu	PgpX11	PgpX11	PgpX13	PgpX13	Pgp24
A	1241219	338755	a01	F	a02	179	185	192	194	0
A	1241313	338737	a02	F	a15	179	179	192	192	0
A	1241313	338737	a02	M	a26	179	179	194	194	197
A	1241401	338702	a03	F	a18	209	209	192	192	197
A	1241401	338702	a03	M	a29	0	0	192	192	197
A	1241500	338734	a04	F	a09	179	195	174	186	197
A	1241500	338734	a04	F	a12	185	185	192	192	197
A	1241601	338736	a05	M	a19	179	179	174	174	197
A	1241725	338719	a06	F	a03	0	0	194	194	197
A	1241725	338719	a06	F	a04	199	209	192	194	197
A	1241725	338719	a06	F	a08	185	195	186	192	197
A	1241725	338719	a06	M	a20	185	185	196	196	197
A	1241967	338750	a07	F	a05	0	0	194	194	197
A	1241967	338750	a07	F	a06	179	179	186	192	197
A	1241967	338750	a07	F	a10	179	185	192	192	197
A	1241967	338750	a07	M	a23	179	179	194	194	0

Tableau 23 (suite)

Site	Longitude	Latitude	Piège	Sexe	Individu	PgpX11	PgpX11	PgpX13	PgpX13	Pgp24
A	1242142	338839	a08	M	a21	185	185	192	192	197
A	1242142	338839	a08	M	a22	195	195	0	0	197
A	1242142	338839	a08	M	a25	185	185	192	192	197
A	1242142	338839	a08	M	a31	185	185	192	192	0
A	1242327	338769	a09	M	a24	185	185	192	192	0
A	1242327	338769	a09	M	a27	185	185	192	192	0
A	1242397	338757	a10	M	a32	185	185	192	192	197
A	1242569	338727	a11	F	a11	179	185	192	194	0
A	1242569	338727	a11	F	a13	179	187	186	192	197
A	1242980	338613	a12	F	a01	0	0	194	198	197
A	1242980	338613	a12	F	a07	181	195	186	192	197
A	1242980	338613	a12	F	a17	181	181	192	192	197
A	1243473	338374	a13	M	a30	0	0	166	166	219
A	1243714	338251	a14	M	a28	197	197	0	0	197
H	1295085	375581	h01	F	h13	0	0	194	194	197
H	1295155	375668	h02	F	h15	185	199	0	0	0
H	1295197	375837	h03	F	h14	179	185	174	186	0

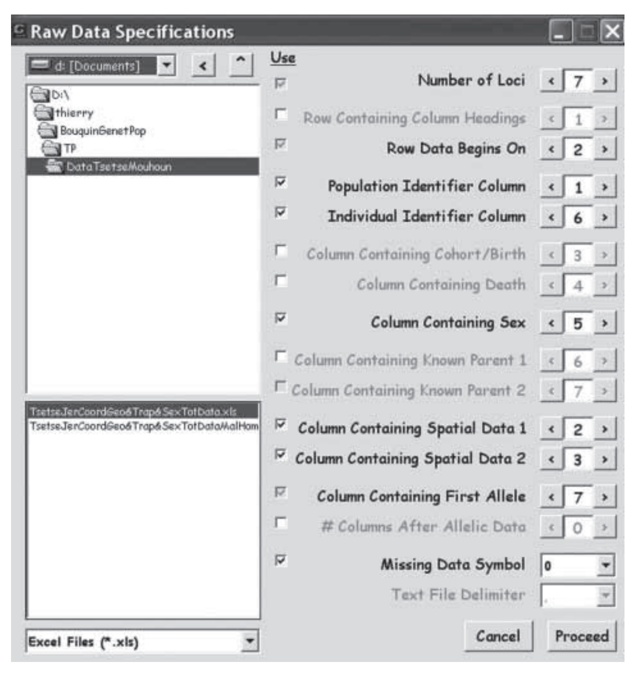


Figure 53
Fiche Menu pour Create pour convertir le fichier de données brutes de mouches tsé-tsé de la Mouhoun au format désiré.

Output Selections

Select Programs For File Output

Sibship Programs: COLONY, KINGROUP, PARENTAGE, PEDIGREE, PRT

Parentage Programs: CERVUS, FAMOZ, FAMSPHERE, GERUD2, GIMLET, NEWPAT, PAPA, PARENTE, PASOS, PEDAPP, PROBMAX3, WHICHPARENTS

Multi-Purpose Genetic Programs: ARLEQUIN, FSTAT, GDA, GENEPOP, GENETIX, MICROSAT, MSA, PHYLIP, SPAGEDI, TFPGA

Specialized Genetic Programs: BAPS4, BATWING, BAYESASS, IM/IMa, IMMANC, LAMARC, MIGRATE, MSVAR, STRUCTURE, COLONISE, FDIST2, NEWHYBRIDS, BOTTLENECK, GENECLASS2, WHICHRUN, CONE, MLNE, NEESTIMATOR, SALMONNB, TM3, TMVP, MICROCHECKER, IDENTIX, ML-RELATE, WHICHLOCI

Deselect All · Select All

Output Folder: D:\thierry\BouquinGenetPop\TP\DataTsetseMouhoun · Change

Output Filename: TsetseMouhouMalHomo

<-Back · CREATE

Figure 54
Second menu de Create pour convertir les données en format Fstat et Genepop.

et qui risquent de générer des problèmes ensuite. Ne gardez que la première colonne de ces fichiers, qui correspond à l'identifiant des sous-populations.

PREMIÈRES ANALYSES : INDÉPENDANCE ENTRE ALLÈLES DANS ET ENTRE LOCI

Déséquilibres de liaison au sein des quatre zones

Lancez Fstat et chargez le fichier “TsetseMouhouMalHomo”. Testez les déséquilibres de liaison en demandant le test “for each pair of loci in each population” et au “nominal level” 1/100 afin d'avoir assez de précision. Dans le fichier de sortie correspondant, nous constatons que seul un test est significatif entre les loci 1 et 2 (c'est-à-dire entre PgpX11 et PgpX13) avec une *P*-value = 0,0044. Cette *P*-value ne reste pas significative après correction de Bonferroni (0,0044×21 = 0,09) et un test significatif sur 21 représente environ 5 % des tests, ce qui est la proportion attendue sous l'hypothèse nulle. Avec la procédure "binom.test" sous R, nous pouvons calculer la probabilité avec laquelle nous pouvons observer une fois un test significatif au seuil α = 0,0044 sous l'hypothèse nulle H_0. Cette probabilité est *P*-value = 0,0889. On peut donc considérer qu'à l'échelle de chaque zone, il y a indépendance entre loci.

Test de la panmixie dans les quatre zones d'échantillonnage

Chargez “DataTsetseMouhounMalManq.dat” dans Fstat et demandez le F_{IS} par locus et population, les estimations de Weir et Cockerham et testez Hardy-Weinberg dans les sous-échantillons avec 10 000 permutations d'allèles entre individus. Le résultat peut être résumé dans le tableau 24. On y constate un fort déficit en hétérozygotes très significatif, mais aussi une forte variance du F_{IS} entre loci. Une recherche d'allèles nuls, de « stuttering » ou de dominance d'allèles courts s'avère nécessaire.

ANALYSE PAR MICRO-CHECKER

Il faut ici traiter les femelles seules pour les loci hétérosomaux du site A (un fichier de plus), pour les autres sites il n'y pas de problème (pas de mâle). En passant par

Tableau 24
Résultat du test de Hardy-Weinberg sur le F_{IS} dans les différentes zones de capture des tsé-tsé, par locus et sur l'ensemble et résultat des tests de permutation.

	Zones				Toutes les zones	*P*-value
Loci	**A**	**H**	**C**	**D**		
PgpX11	0,253	0,258	0,239	0,105	0,220	0,0003
PgpX13	0,137	0,131	0,097	0,251	0,157	0,0055
Pgp24	0,662	0,375	0,086	0,339	0,271	0,0001
B11_1	0,194	0,189	0,305	0,344	0,262	0,0001
BX104	0,269	0,436	0,086	0,097	0,214	0,0005
C102	0,125	- 0,058	- 0,133	0,499	0,137	0,0874
GpCag	- 0,058	- 0,13	- 0,074	0,068	- 0,052	0,8244
Tous les loci	0,175	0,199	0,105	0,235	0,175	0,0001

Create, vous transformez vos deux fichiers Excel en format Micro-Checker, le premier pour les loci liés à l'X en A, le second pour toutes les données (et on ne regardera pas le résultat des loci liés au sexe en A). Lancez micro-Checker. N'oubliez pas de préciser le pas de mutation correct. Bon je vous aide, mis à part BX104 et GpCag (mononucléotides) et C102 (trinucléotides), tous les loci sont dinucléotidiques. Les analyses montrent que les allèlles nuls expliquent très bien tous les résultats, y compris pour les mâles. En effet, pour les femelles et loci autosomaux de A, et pour tous les loci en H, C et D, il y a plus de blancs observés qu'attendus par la méthode de Brookfield. Pour les loci liés à l'X chez les mâles du site A, les différences ne sont pas significatives. Il semble même y avoir du « stuttering » pour le locus C102 en D. Cependant, l'effet Wahlund ne peut non plus être totalement écarté, ainsi que nous allons pouvoir le vérifier. Vous pourrez aussi vérifier qu'il ne semble pas exister de dominance d'allèles courts ici.

MISE EN ÉVIDENCE D'UNE SOUS-STRUCTURATION À L'INTÉRIEUR DES ZONES A, H, C ET D

Vous pouvez tester en zone A s'il existe un biais de structuration spécifique au sexe entre pièges, à titre d'exercice, et constater qu'il n'y a aucune signature d'un tel

phénomène dans ces données. Nous allons rechercher un possible effet Wahlund comme une cause possible d'excès d'homozygotie chez les tsét-tsé d'une même zone : d'abord en analysant le F_{IS} à une échelle plus réduite (piège), ensuite par analyse bayésienne de clusterisation comme pour les tiques et enfin en recherchant un isolement par la distance entre individus le long du cours d'eau.

Analyse par piège

En prenant chaque piège comme une sous-population potentielle et en recalculant le F_{IS}, on obtient une valeur plus faible de 0,144, significativement inférieure à la précédente (test de Wilcoxon pour données appariées comme pour les tiques, *P*-value = 0,0391), mais toujours significativement supérieure à 0 (*P*-value = 0,0001). Il semble donc bien que chaque piège recèle, au moins en partie, des mouches plus apparentées que des mouches prises au hasard dans chaque zone. À cause de la faiblesse des échantillons, Micro-Checker ne peut être utilisé ici. Nous devons donc trouver une méthode alternative afin de rechercher si les allèles nuls peuvent contribuer à expliquer les déficits en hétérozygotes rencontrés. Une méthode pratique consiste à regarder s'il existe une relation entre le nombre de blancs par locus et le F_{IS} effectivement mesuré à ce locus, dans chaque sous-échantillon. Nous obtenons ainsi les données du tableau 25.

On lance ensuite une analyse de corrélation. Pour plus de sécurité, on utilisera une analyse dite non paramétrique à l'aide du coefficient de corrélation de Spearman. Dans R, les commandes seront (en respectant les majuscules et minuscules, test unilatéral car on a un préjugé de la direction du signal) :

```
> data<-read.table("BlancFisTsetse.txt",header=TRUE)
> attach(data)
> cor.test(data$Blancs, data$Fis, alternative="greater",
method="spearman")
```

Le résultat est un coefficient de corrélation de Spearman ρ = 0,46 très significatif (*P*-value = 0,0073) (fig. 55).

Vous remarquerez que le coefficient de corrélation est légèrement différent de celui publié dans l'article de *Molecular Ecology* (ρ = 0,499 et *P*-value = 0,0048). La différence provient de trois pièges de l'échantillon en zone A (le seul où il y avait des mâles) où le F_{IS} est différent. Cela provient certainement du recodage des mâles ou plus probablement du fait que je n'ai éliminé aucun sous-échantillon ici, même ceux de taille 1. De toutes manières, cela ne change pratiquement rien. Ce genre de petits problèmes est fréquent quand le nombre d'analyses différentes à effectuer est très grand, comme cela a été le cas ici. C'est pour cela que j'ai choisi d'en parler, car cela arrive et il ne faut pas le cacher. Ce genre d'erreurs (assimilables à celles

Tableau 25
F_{IS} et nombre de blancs (homozygotes nuls supposés) par piège-site et par locus.

Site	Locus	Blancs	F_{IS}
A	PgpX11	5	0,13
A	PgpX13	2	0,299
A	Pgp24	7	0,165
A	B11	7	0,218
A	BX104	4	0,1
A	C102	6	0,125
A	GpCag	1	- 0,274
H	PgpX11	6	0,288
H	PgpX13	2	- 0,125
H	Pgp24	7	0,557
H	B11	4	0,174
H	BX104	4	0,444
H	C102	7	- 0,197
H	GpCag	1	- 0,247
C	PgpX11	9	0,235
C	PgpX13	0	0,024
C	Pgp24	3	- 0,088
C	B11	2	0,193
C	BX104	1	0,066
C	C102	2	- 0,043
C	GpCag	1	0,079
D	PgpX11	12	0,214
D	PgpX13	1	0,22
D	Pgp24	3	0,175
D	B11	3	0,207
D	BX104	6	0,185
D	C102	6	0,25
D	GpCag	4	0,148

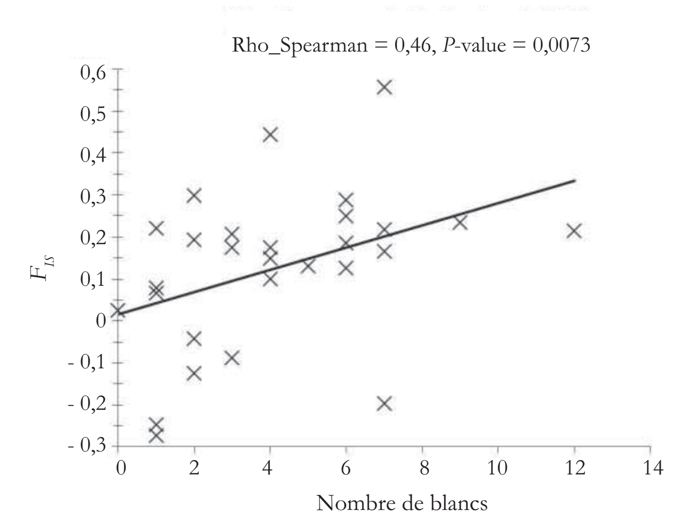


Figure 55
Corrélation entre nombre de blancs observés et valeur du F_{IS} par locus et piège-site.

éventuelles associées au génotypage/sexage, etc.), inévitables à la longue, n'est cependant pas en mesure de générer un signal quelconque, et va plutôt contribuer à masquer les signaux de faibles amplitudes. Ici, les allèles nuls expliquent donc bien en partie les F_{IS}. En mettant au carré le coefficient de corrélation trouvé, on réalise qu'environ 21 % seulement de la variance de ces derniers est expliquée par ce phénomène (16 % si on utilise le coefficient de détermination normal). Il est donc raisonnable de rechercher d'autres facteurs responsables de ces déficits en hétérozygotes.

Clusters BAPS

Ici, en ce qui concerne mon analyse, les clusters trouvés par BAPS semblent expliquer une très grande partie du déficit en hétérozygotes, voire la totalité, puisque nous passons d'un F_{IS} = 0,175 à un F_{IS} = 0,031 non significativement différent de 0 cette fois (~ panmixie locale). Il semble donc bien que l'effet Wahlund soit responsable de la plus grande part du déficit en hétérozygotes. Il semble aussi que les pièges eux-mêmes capturent des mouches issues de voisinages différents puisque le F_{IS} intra-piège, même s'il baisse, reste fortement positif. La correspondance entre les clusters BAPS et pièges est à cet égard mauvaise (vérifiez-le), même s'il arrive fréquemment que des mouches du même piège se retrouvent dans le même cluster BAPS. Les allèles nuls jouent peut-être, quant à eux, un petit rôle également, comme semblent le montrer les analyses de MicroChecker et de corrélation avec le nombre de blancs. Cependant, une régression du nombre de blancs trouvés dans les

différentes zones n'explique que peu la dispersion des F_{IS} par loci et piège-zone (R^2 = 0,16). Cependant, en cas de pangamie, c'est un F_{IS} légèrement négatif qui est attendu. Le F_{IS} des clusters de BAPS étant légèrement positif, il est possible que la contribution des allèles nuls, même modeste, soit réelle. Mais c'est bien l'effet Wahlund qui explique le mieux les données.

Isolement par la distance entre individus

La plupart de ces pièges contiennent trop peu d'individus génotypés pour mettre en œuvre la même procédure que pour les tiques. Par ailleurs, nous savons que les pièges ne représentent qu'approximativement des voisinages (si voisinage il y a) puisque pièges et clusters BAPS ne sont pas en très bon accord. Nous pouvons cependant utiliser ici la procédure d'isolement par la distance entre individus (Rousset, 2000 ; Watts *et al.*, 2007) implémentée par le logiciel Genepop 4 (Rousset, 2008) téléchargeable à partir du site http://kimura.univ-montp2.fr/~rousset/Genepop.htm. Il faut recoder les données pour chaque zone (un fichier par zone A, H, C et D) de telle sorte que chaque individu est considéré comme une sous-population comme dans la figure 56.

Il faut ensuite copier le logiciel Genepop.exe dans le répertoire où se trouvent les quatre fichiers que nous venons de créer. En ce qui me concerne, et n'écoutant que mon imagination débordante, j'ai nommé les quatre fichiers A.txt, H.txt, C.txt et D.txt. Ici, si vous êtes cardiaque, je vous conseille de couper le son de votre ordinateur, car quand Genepop est content il le manifeste par un bruit effroyable à réveiller un mort. On clique deux fois sur Genepop.exe et une fenêtre apparaît où le nom du fichier vous est demandé. À l'invite, tapez "A.txt" puis "Entrée". Lisez les informations et si vous êtes d'accord retapez "Entrée". Un menu apparaît. C'est l'option 6 qui nous intéresse. Tapez donc "6". Un sous-menu apparaît dont l'option 5 est celle qu'il faut implémenter. Tapez "5". On vous demande si vous souhaitez effectuer le test avec la statistique $\hat{a}$ (un équivalent du $F_{ST}/(1 - F_{ST})$ pour la différenciation entre individus) ou $\hat{e}$. D'après Watts *et al.* (2007), pour une structure en une dimension, comme c'est le cas le long de la rivière Mouhoun, la statistique $\hat{e}$ est meilleure quand la taille de voisinage (*Nb* pour *neighbourhood*) $Nb = 4D\sigma^2 > 10\,000$ individus et $\hat{a}$ est plus performant quand $Nb < 10\,000$ individus. Commençons par $\hat{a}$ et nous prendrons $\hat{e}$ ensuite. Tapez "a". On vous demande si vous souhaitez faire le test avec le logarithme népérien des distances géographiques ou non. Tapez "d" car nous sommes dans un contexte unidimensionnel (cf. p. 89 en première partie). On vous demande la distance minimale à considérer pour la régression. Comme le test n'en tiendra pas compte, que le biais ne risque pas d'être important (Watts *et al.*, 2007, voir aussi le commentaire de Rousset dans la documentation de Genepop 4) et qu'il n'y a pas de log, tapez "0". Le nombre de randomisations à effectuer pour le test de Mantel vous est demandé. Tapez "1 000 000". En

```
Data·tsetse·GPS·Mouhoun·zone·A¶
PgpX11¶
PgpX13¶
Pgp24¶
B11¶
BX104¶
C102¶
GpCag¶
pop¶
1241219»        338755,»    179185»    192194»    000000»    169
pop¶
1241313»        338737,»    179179»    192192»    000000»    169
pop¶
1241313»        338737,»    179179»    194194»    197197»    195
pop¶
1241401»        338702,»    209209»    192192»    197197»    169
pop¶
1241401»        338702,»    000000»    192192»    197197»    000
pop¶
1241500»        338734,»    179195»    174186»    197197»    000
pop¶
1241500»        338734,»    185185»    192192»    197197»    169
pop¶
1241601»        338736,»    179179»    174174»    197197»    173
pop¶
1241725»        338719,»    000000»    194194»    197197»    169
pop¶
1241725»        338719,»    199209»    192194»    197197»    167
pop¶
1241725»        338719,»    185195»    186192»    197197»    000
pop¶
1241725»        338719,»    185185»    196196»    197197»    169
pop¶
1241967»        338750,»    000000»    194194»    197197»    169
pop¶
1241967»        338750,»    179179»    186192»    197197»    183
pop¶
1241967»        338750,»    179185»    192192»    197197»    169
pop¶
1241967»        338750,»    179179»    194194»    000000»    169
```

Figure 56
Extrait du fichier des données recodées pour le site A avec les coordonnées GPS des pièges pour l'analyse d'isolement par la distance entre individus. Chaque individu est séparé des autres par un "pop" et codé par sa longitude puis latitude une « , » et les génotypes aux loci microsatellites. Les mâles (seulement en A) sont codés homozygotes pour les loci liés à l'X (indiqué par un X dans le nom du locus).

fonction de l'ordinateur le processus Markovien prend plus ou moins de temps. Le programme vous demande te taper "Return" (soit "Entrée"). Le résultat est disponible dans A.txt.ISO. On fait de même avec H, C et D. Pour changer de fichier de données, il faut taper "C" dans le menu général de Genepop. Ne soyez pas étonnés si, à partir de H, le test de Mantel démarre sans vous demander votre avis. C'est comme ça. C'est Genepop. Vous vous apercevez que le calcul ne se fait pas pour D. En fait, cela ne se termine jamais, car il y a un problème dans le fichier et un bug dans Genepop. Ouvrez D.txt. Il faut supprimer le 12[e] individu (000000 partout), sauvez puis recommencez, ça marche ! Ensuite, on enregistre les quatre fichiers de données sous un autre nom pour les analyses avec *ê*, par exemple A_e.txt, H_e.txt, C_e.txt et D_e.txt (quelle imagination ! mais où va-t-il les chercher ?). Nous nous retrouvons donc avec huit fichiers *.ISO que

nous pouvons ouvrir avec n'importe quel éditeur de texte. Vous pouvez aussi remettre le son. Nous souhaitons savoir si $4D\sigma^2 > 10\ 000$. D'après ce que nous avons vu en p. 89 de la première partie de ce manuel, le voisinage est égal à $Nb = 1/b = 4D\sigma^2$. Nous souhaitons vérifier si $Nb > 10\ 000$ afin de décider si c'est le paramètre $\hat{a}$ ou $\hat{e}$ qu'il vaut mieux utiliser. C'est le cas uniquement pour le site A avec la statistique $\hat{a}$. En outre, vous remarquerez que la statistique $\hat{a}$ donne de bien meilleurs résultats avec ces données de glossines. On sait par ailleurs que ce type de tests est très conservateur et que $\hat{a}$ n'est pas biaisé alors que $\hat{e}$ l'est (WATTS *et al.*, 2007). Nous ne considérerons donc que les résultats obtenus avec $\hat{a}$.

Les résultats pour les quatre zones et la moyenne sur l'ensemble figurent dans le tableau 26. Il y a donc bien un isolement par la distance, mais les pentes sont très faibles. Cela signifie que les voisinages sont très lâches (beaucoup d'échange entre voisins, σ grand) et/ou de grande taille (D grand). Pour visualiser cette relation, nous pouvons utiliser les sorties *.GRA de genepop qui contiennent deux colonnes, la première avec les distances géographiques et la seconde avec la distance génétique *a*. Nous pouvons les charger sous Excel en précisant que les colonnes sont délimitées par des espaces et tracer le graphique de la figure 57.

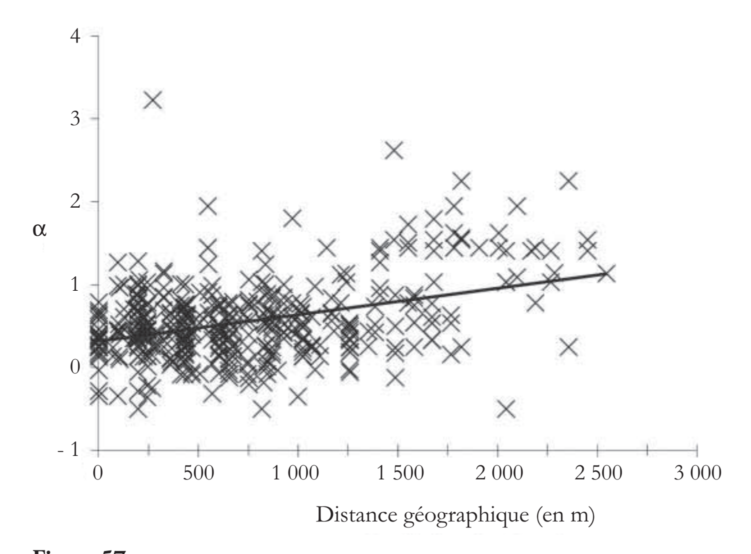


Figure 57
Représentation graphique de l'isolement par la distance entre individus des mouches tsé-tsé le long du Mouhoun en zone A.

Nous avons maintenant besoin d'estimer des effectifs efficaces pour essayer d'obtenir une idée (mais ce sera à la louche) des densités.

Tableau 26
Résultats de l'analyse d'isolement par la distance entre individus pour les quatre sites (A, H, C, D) le long de la rivière Mouhoun au Burkina Faso. La pente *b* de la régression, la taille efficace de voisinage *Nb*, le produit de la densité efficace par la surface efficace de dispersion $D\sigma^2$ et la *P*-value du test sont donnés, ainsi que les moyennes non pondérées pour *b*, *Nb* et $D\sigma^2$. Les *P*-value ont été combinées par la méthode binomiale généralisée avec MultiTest.

	b	*Nb*	$D\sigma^2$	*P*-value
A	0,000322	3105	776	0,0055
H	8,02E-06	124725	31181	0,3805
C	6,26E-06	159755	39939	0,2056
D	8,22E-06	121713	30429	0,0237
Moyenne	8,61E-05	102325	25581	0,0033

Effectifs efficaces

Les seules méthodes disponibles ici sont celles basées sur l'hétérozygotie et les déséquilibres de liaison, où nous serons obligés de considérer l'absence d'allèles nuls et d'effet Wahlund. Ces phénomènes (que nous savons probables pour ces échantillons de mouches tsé-tsé) auront tendance à produire des surestimations de N_e pour les méthodes basées sur l'hétérozygotie, et des sous-estimations pour les méthodes basées sur les déséquilibres de liaison. Nous utiliserons trois méthodes. La méthode d'identité intra et inter locus de Vitalis et Couvet (Vitalis et Couvet, 2001a, b, c) est implémentée par le logiciel Estim qui accepte le format Genepop pour le fichier de données. La méthode des déséquilibres de liaison de Bartley *et al.* (1992) est modifiée comme décrit dans l'aide du logiciel NeEstimator. La méthode des excès en hétérozygotes se fait très simplement en utilisant l'estimateur de Weir et Cockerham du F_{IS} dans la formule $N_e = 1/(-2F_{IS}) - F_{IS}/(1 + F_{IS})$ (Balloux, 2004) qui ne donne bien entendu un résultat valide que si $F_{IS} < 0$.

Pour ce faire, les effectifs par piège étant bien insuffisants, nous allons devoir regrouper les mouches de différents pièges en fonction de leur proximité (voir fig. 52). Ceci ne va pas arranger l'effet Wahlund, mais nous n'avons pas le choix. Ces regroupements sont synthétisés dans le tableau 27 (trois premières colonnes). Les pièges isolés ne contenant qu'une seule mouche ne sont pas représentés dans ce tableau. Vous pouvez essayer avec une autre stratégie de regroupement pour vérifier si on retrouve des valeurs équivalentes. Pour Estim, il faut recoder les données de départ au format Genepop, avec données manquantes pour les mâles aux loci hétérosomaux, et les regroupements du tableau 27. Pour NeEstimator

(déséquilibres de liaison), il faut autant de fichiers qu'il y a de groupes de pièges définis dans le tableau 27 avec les mâles codés homozygotes pour les loci liés à l'X. Les même fichiers que pour Estim pourront être utilisés pour l'estimation des F_{IS} par groupe de pièges, soit en utilisant Genepop directement, soit en traduisant les fichiers pour un autre logiciel (Fstat, Genetix). Notez que NeEstimator donne aussi une estimation basée sur les excès d'hétérozygotes (Luikart et Cornuet, 1999), mais contenant des inexactitudes corrigées par la méthode de Balloux. On peut aussi utiliser le fichier recodé Genepop pour une analyse par LDNe (Waples et Do, 2008), qui implémente une méthode basée sur les déséquilibres de liaison non biaisée (ou beaucoup moins) pour les petits échantillons, alors qu'on sait que la méthode de Bartley est biaisée quand la taille des échantillons est inférieure à la taille efficace des populations étudiées (England *et al.*, 2006 ; Waples, 2006). Ceux qui s'en rappellent constateront que les méthodes implémentées par Estim et celles basées sur les excès d'hétérozygotes n'avaient pas été utilisées pour les tiques (p. 160-164 de la seconde partie). Chez les tiques, la forte présence d'allèles nuls en plus de la dominance des allèles courts au locus IR27 rendaient caduque toute approche basée sur les corrélations d'allèles intra-individuelles. Ici, il n'y a pas de dominance d'allèles courts et les allèles nuls sont peu influents, même si on ne peut totalement exclure leur impact (voir plus haut).

Tableau 27
Stratégie de regroupements par piège de Gpg le long du Mouhoun, en se basant sur la figure 52 et estimation des effectifs efficaces. Les résultats sont donnés pour les trois méthodes utilisées pour des résultats autres que l'infini, 0 ou NA (not available) (cases vides). Le nombre d'individus génotypés par piège est donné (N_{Traps}).

Zone	Pièges	N_{Traps}	Estim	Déséquilibres de liaison	Excès d'hétérozygotes
A	1, 2, 3	1, 2, 2		1,6	
A	4, 5, 6	2, 1, 4		5	
A	7	4	3,19		
A	8	4			3,4
A	9, 10, 11	2, 1, 2		1,3	4,7
A	12	3		0,3	
A	13, 14	1, 1		0,7	
H	1, 2	1, 1			
H	3, 4	1, 2			25

Tableau 27 (suite)

Zone	Pièges	N_{Traps}	Estim	Déséquilibres de liaison	Excès d'hétérozygotes
H	5	2			
H	6, 7	1, 3			
H	8	3			
H	9, 10, 11, 12	1, 1, 3, 2			
H	14, 15	1, 3			
H	16, 17	1, 2	2,08		
C	1	3			
C	2, 3, 4	1, 1, 1			
C	5, 6	1, 2			
C	7, 8	1, 2			3,8
C	10, 11	1, 1			
C	12	2			
C	13	2			
C	14, 15	1, 2			
C	19	3			
D	2, 3, 4	1, 3, 2		2,2	
D	6	3		2,4	
D	8	3		1,5	
D	11	2			7,5
D	12	2			
D	13	3		0,6	
D	14	3		0,4	
D	16	2			

Le logiciel Estim (http://www.ecoanthropologie.cnrs.fr/spip.php?article296) utilise un fichier au format Genepop. Dans la mesure où Estim utilise les identités intra-individuelles, interindividuelles, inter-échantillons et leur corrélation entre loci, et que par ailleurs l'hypothèse d'un modèle en îles est faite, il est clair que nous ne sommes pas

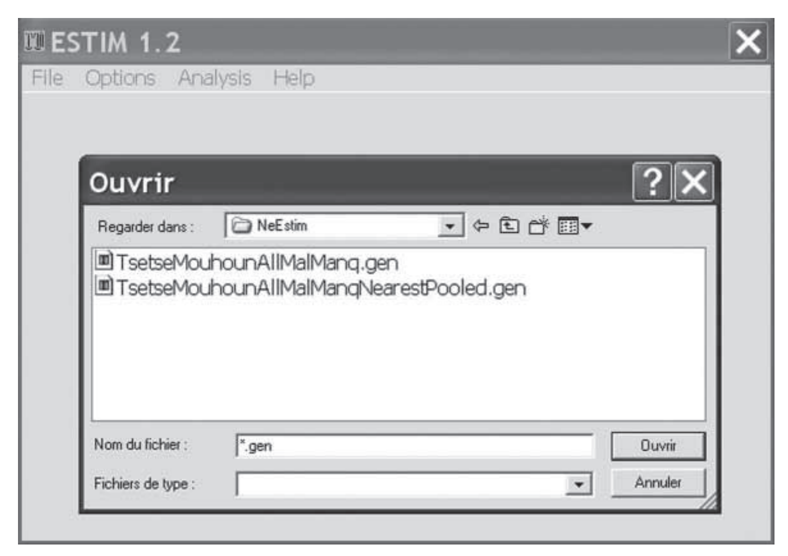


Figure 58
Chargement des données dans le logiciel Estim, pour estimation des N_e à partir des déséquilibres intra et inter-loci (données manquantes chez les mâles pour les loci liés à l'X).

tout à fait dans les critères orthodoxes de cette méthode. Il vaut mieux considérer chaque zone (A, H, C, D) séparément, car cela influence les résultats (comme vous pourrez le vérifier). Notez que la stratégie de regroupement diffère quelque peu de celle de l'article de Bouyer *et al.* (2009) avec des résultats légèrement différents. On peut donc charger le fichier contenant tous les groupes de pièges de la zone A dans Estim. Mon fichier s'appelle "TsetseMouhounAllMalManqNearestPooledA.gen". N'oubliez pas de supprimer les pièges isolés ne contenant qu'une mouche, car Estim ne va pas apprécier. On lance donc Estim et on charge son fichier. Pour qu'il apparaisse, on tape *.gen dans la case appropriée comme dans la figure 58 ou alors on change l'extension du fichier de .gen à .txt et on clique dessus deux fois.

Ensuite, on appelle la commande "Identity measures" du menu "Analysis" (fig. 59).

On obtient alors une fenêtre résultat dont on fait descendre le curseur pour pouvoir enregistrer (cliquer sur "Save") (fig. 60). Je l'ai enregistrée sous le nom "NeEstimA.txt"

Cliquez ensuite sur l'option "Ne inferences" du menu "Analysis" et sauver en gardant le même nom, car ces nouveaux résultats sont écrits à la fin de la fenêtre précédente. Vous pouvez ouvrir le fichier résultat avec un éditeur de texte. Recommencez la même opération pour chacune des zones restantes. Les résultats sont que seules deux collections de pièges donnent des valeurs exploitables : le piège 8 de la zone A (N_e = 3,19, m = 0,27) et le groupe de pièges (16, 17) de la zone H (N_e = 2,08, m = 0,55) (tabl. 27).

La méthode de Waples et Do, pour laquelle il suffit de charger le fichier Genepop avec toutes les données en appuyant sur le bouton "Search", ne donne aucun résultat



Figure 59
Menu pour lancer la première analyse à effectuer avec ESTIM.

ici (pas d'estimation possible) comme c'est très souvent le cas, mais la plupart des limites inférieures paramétriques disponibles indiquent de très faibles valeurs de N_e. Néanmoins, et sans garde-fou solide pour la méthode implémentée, il faudra bien garder à l'esprit qu'on aura peut-être ici des valeurs très sous-estimées par la méthode des déséquilibres de liaison de Bartley. La méthode des déséquilibres de liaison de Bartley est implémentée par NeEstimator comme pour les tiques (un fichier par groupe de piège ici) (voir p. 160-164 dans la seconde partie de ce manuel). Enfin,

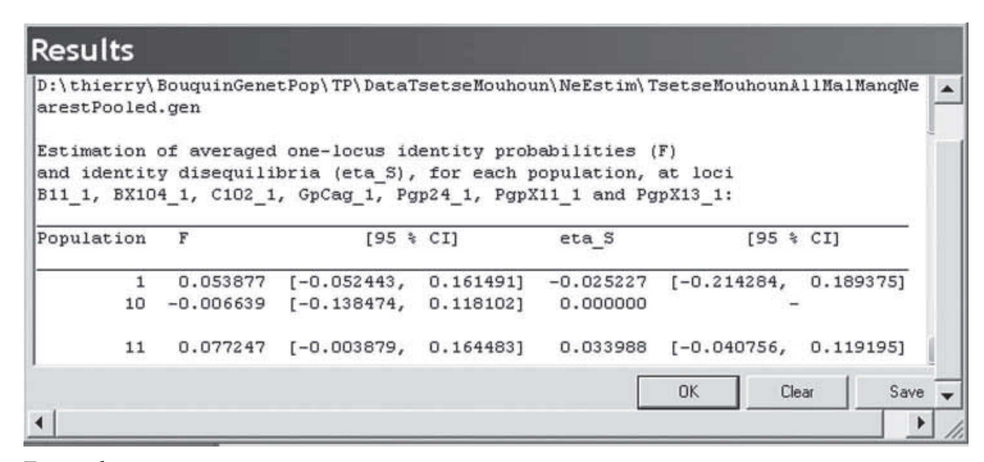


Figure 60
Cadre de première sortie et de création du fichier de sauvegarde de l'analyse par ESTIM.

la méthode de Balloux est très simple puisqu'il suffit de calculer les F_{IS} par groupe de pièges (avec Fstat, par exemple) et d'utiliser la formule $N_e = 1/(-2F_{IS}) - F_{IS}/(1 + F_{IS})$ et de ne garder que les valeurs de N_e positives. Tous les résultats sont compilés dans le tableau 27. Nous avons maintenant besoin de transformer ces effectifs en densités.

Densités efficaces

Nous allons utiliser une autre méthode que celle de Bouyer *et al.* (2009), pour changer. Nous allons simplement pour chaque méthode prendre l'effectif efficace moyen par piège (pondéré pour le nombre d'individus et de pièges) $\overline{N_e}$ et diviser cette valeur par la distance minimale (en m) entre deux pièges (tabl. 28). Cette distance minimale se trouve en zone A (facile à trouver dans le fichier que vous avez créé pour la figure 56) et est de $D_{\text{mini}} = 71$ m. Le calcul de pondération est assez particulier. Pour ce faire, j'ai multiplié le N_e par le nombre d'individus N_{ind} capturés dans les pièges correspondants : $N_{eP} = N_e \times N_{ind}$. Pour chaque N_e, j'ai calculé le produit du nombre de mouches par le nombre de pièges correspondant $n_{pièges}$: $N_P = N_{ind} \times n_{pièges}$. J'ai ensuite fait la somme des $N_{eP} = \sum N_{eP}$ et des $N_P = \sum N_P$. L'effectif efficace moyen est ensuite calculé par le rapport de ces deux valeurs.

$$\overline{N_e} = \frac{\sum N_{eP}}{\sum N_P}$$

La densité D_c est ensuite obtenue en divisant cette valeur par $D_{\text{mini}} = 71$, ce qui permet ensuite de déduire les dispersions σ à partir des valeurs de $D\sigma^2$ du tableau 26 :

$$\sigma = \sqrt{\frac{D\sigma^2}{D_c}}$$

On comprend bien que les valeurs obtenues (tabl. 28) ne pourront être que très approximatives.

Par conséquent, à partir des données génétiques et de leur analyse (isolement par la distance) et de calculs de densités efficaces, nous pouvons inférer que le long du Mouhoun les tsé-tsé ont des densités comprises entre 12 et 176 mouches par km et une dispersion (distance entre adultes reproducteurs et leurs parents) comprise entre 131 m et 1 620 m. Ces inférences sont remarquablement convergentes avec celles du papier de Bouyer *et al.* (2009) (tabl. 29) et donc avec les données issues de marquage-relâchage et recapture (MRR) de tsé-tsé marquées en zone A. Dans la mesure où la stratégie de regroupement fut ici légèrement différente, de même que le choix de calcul des densités, ce résultat confirme la robustesse des résultats du papier. La convergence avec les données de marquage-recapture souligne également l'efficacité des outils de génétique des populations, en particulier la méthode de Rousset (1997) pour estimer $D\sigma^2$.

Tableau 28
Calcul détaillé des densités (en mouches par m) et de la dispersion des glossines (distance entre adultes reproducteurs et leurs parents en m) le long du Mouhoun.

$\overline{N_e}$				
Zone	**Estim**	**Bartley**	**Balloux**	D_{mini} **(m)**
A	3,19	0,928	1,72	71
H	1,04		12,5	
C			1,9	
D		0,823	7,5	
D **(densité efficace)**				
Zone	**Estim**	**Bartley**	**Balloux**	$D\sigma^2$
A	0,045	0,013	0,024	776,27
H	0,015		0,176	31 181,25
C			0,027	39 938,72
D		0,012	0,106	30 428,36
σ **(dispersion)**				
Zone	**Estim**	**Bartley**	**Balloux**	
A	131	244	179	
H	1 459		421	
C			1 222	
D		1 620	537	

Conclusions : isolement par la distance intra-zone *(rolling on the river)*

Nous avons bien mis en évidence une sous-structure au sein des zones A, H, C et D. Le fait que les déficits en hétérozygotes persistent au sein de chaque piège, auquel s'ajoute la non-correspondance parfaite entre clusters BAPS et pièges alors qu'un isolement par la distance existe bel et bien, plaide pour deux interprétations complémentaires. Il semble bien y avoir quelques allèles nuls, mais ces derniers n'expliquent qu'une faible partie des déficits en hétérozygotes observés. L'effet Wahlund explique probablement la majeure partie des déficits. Il provient de deux causes. La première est inhérente aux systèmes d'isolement par la distance, d'une nature plus ou moins continue, et de la nature nécessairement discrète du piégeage des tsé-tsé. La seconde raison, qui dépend de

Tableau 29
Estimation des densités (en mouches par m) et de la dispersion des glossines (en m) le long du Mouhoun et moyennées sur l'ensemble des méthodes (All). Les valeurs correspondantes obtenues par MRR (MRR) sont également fournies (d'après BOUYER *et al.*, 2009).

Site	$D\sigma^2$	D_c	σ
A	776,277	0,033	153
H	31 210,986	0,128	493
C	39 936,102	0,036	1053
D	30 413,625	0,086	596
All	29 02,421	0,071	574
MRR		0,2	[1 245, 2 392]

la première, provient de la mauvaise correspondance entre dispersion trophique, plus large, et dispersion reproductrice (accouplements et larvipositions) plus restreinte (homing). Cette information est capitale si nous parvenons un jour à déterminer avec précision les micro-conditions écologiques qui poussent les tsé-tsé à revenir se reproduire et larviposer à l'endroit où elles ont émergé. Il reste aussi à déterminer quelle influence la densité (compétition) a sur la dispersion de reproduction afin d'évaluer si nos estimations restent valables dans le cadre de campagnes de contrôle et/ou d'élimination.

DIFFÉRENTIATION ENTRE LES QUATRE ZONES

Analyse HierFstat du jeu de données total partitionné par BAPS

Nous savons qu'une différenciation existe bien à une mini (voire micro) échelle à l'intérieur de chacune des zones A, H, C et D. Nous devons donc tenir compte de ce niveau de structuration à micro-échelle avant d'estimer et tester l'existence d'une différenciation entre zones. Cette information est utile, car elle pourrait permettre d'estimer le temps nécessaire à une recolonisation d'une zone éliminée par la zone la plus proche.

Nous allons devoir utiliser HierFstat une nouvelle fois. Considérant que les regroupements définis par BAPS pourraient mieux regrouper les individus de la même unité populationnelle par rapport aux pièges et pour faire autre chose que dans le papier initial, nous allons prendre comme niveau le plus imbriqué les clusters BAPS de chaque zone définis en p. 199. Le niveau suivant sera la zone (A, H, C, D) et enfin la totalité. Il y aura ainsi quatre niveaux définis avec leur F, l'individu (F_{IS}), le

sous-groupe défini par BAPS dans la zone (F_{SZ}), la zone dans le tout (F_{ZT}), auxquels s'ajoutent bien sûr les F_{IZ}, F_{IT} et F_{ST} moins intéressants pour nous.

En procédant comme pour les tiques (voir p. 156 dans la seconde partie de ce manuel), et en prenant soin de recoder les mâles homozygotes pour les locus hétérosomiques, on obtient une forte valeur pour $F_{SZ} \approx 0{,}22$ et une valeur négative pour $F_{ZT} \approx$ - 0,03. Il semble que toute l'information soit contenue à l'intérieur des zones et qu'il ne reste plus assez de variation pour distinguer les zones entre elles. Le « *supplementary information* » de ROUGERON *et al.* (2009), présenté ci-dessous, permet de mieux comprendre ce problème inhérent aux statistiques *F* hiérarchiques et renforcé par l'homoplasie des microsatellites.

Comprendre le manque de structure inter-zones avec un peu de théorie

Quand nous avons, comme c'est le cas ici, quatre niveaux hiérarchiques (individus, sous-populations, archipels et totalité), quatre paramètres d'identité peuvent être définis : Q_i, la probabilité que deux allèles d'un locus d'un individu pris au

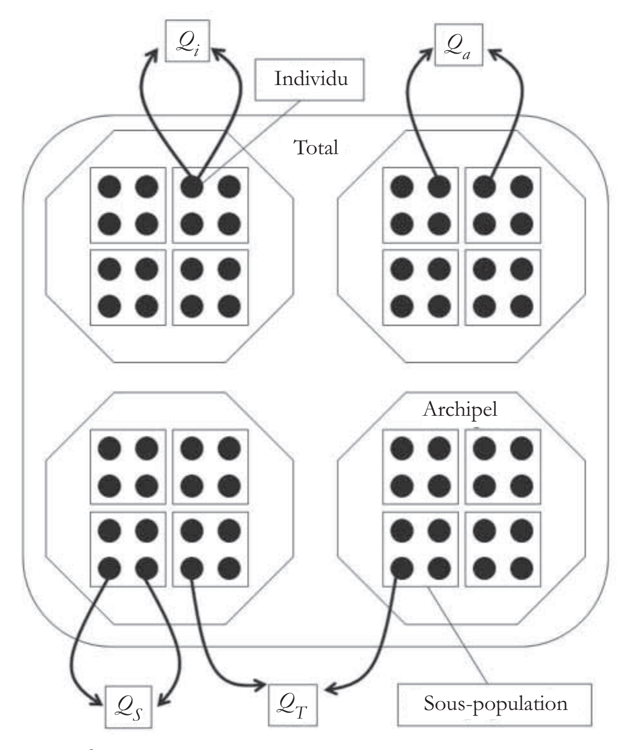


Figure 61
Représentation schématique d'une population structurée en quatre niveaux hiérarchiques, individu, sous-population, archipel et population totale (d'après ROUGERON *et al.*, 2009).

hasard soient identiques ; Q_s, la probabilité que deux allèles à un locus, de deux individus pris au hasard dans la même sous-population soient identiques ; Q_a, la probabilité que deux allèles à un locus, de deux individus pris au hasard dans deux sous-populations différentes dans un même archipel soient identiques ; et Q_T, la probabilité que deux allèles à un locus, de deux individus pris au hasard dans deux sous-populations différentes et deux archipels différents soient identiques (cf. fig. 61).

Nous pouvons définir six indices de fixation : F_{IS} (consanguinité individuelle relative à celle des sous-populations), F_{SA} (consanguinité des sous-populations relative à celle des archipels), $F_{IA} = 1-(1 - F_{IS})(1 - F_{SA})$ (consanguinité individuelle relative à celle des archipels), F_{AT} (consanguinité des archipels relative à celle de la population totale), $F_{ST} = 1-(1 - F_{SA})(1 - F_{AT})$ (consanguinité des sous-populations relative au total) et $F_{IT} = 1-(1 - F_{IS})(1 - F_{ST})$ (consanguinité des individus relative à la population totale). Ces indices peuvent être exprimés, en suivant la méthode proposée par Cockerham (1969, 1973), en fonction des probabilités d'identité définies plus haut dans ce paragraphe (on peut aussi consulter les p. 40-48 de la première partie de ce manuel) :

$$\begin{cases} F_{IS} = \dfrac{Q_i - Q_s}{1 - Q_s} \\ F_{SA} = \dfrac{Q_s - Q_a}{1 - Q_a} \\ F_{IA} = \dfrac{Q_i - Q_a}{1 - Q_a} \\ F_{AT} = \dfrac{Q_a - Q_T}{1 - Q_T} \\ F_{ST} = \dfrac{Q_s - Q_T}{1 - Q_T} \\ F_{IT} = \dfrac{Q_i - Q_T}{1 - Q_T} \end{cases} \qquad (66)$$

Si nous nous concentrons maintenant sur les indices de fixation qui reflètent la différenciation génétique entre sous-populations du même archipel et entre archipels, il n'y a alors plus que F_{SA} et F_{AT} qui nous intéressent. Si pour une raison quelconque, Q_a est très petit (migration très faible entre sous-populations), il est alors facile de voir par l'équation (66) que F_{SA} sera très grand ($\sim Q_s$ si $Q_a \sim 0$). Si la migration est très faible entre sous-populations d'un même archipel, il est alors probable que celle entre archipels soit au moins aussi faible et donc que $Q_T \sim 0$ et $F_{AT} \sim Q_a$. À partir de là, il est facile de voir que, quand la différenciation est extrême entre les sous-populations celle-ci sera nécessairement faible (en apparence) entre archipels. En

fait, cela veut juste dire que la différenciation entre sous-populations est très forte, que ce soit entre sous-populations du même archipel ou de deux archipels différents, et la distinction entre archipels n'apporte pas suffisamment d'information avec ces outils. Ce dernier point peut être illustré mathématiquement par le fait que dans ce cas $F_{ST} \sim F_{SA}$.

C'est donc probablement ce phénomène qui empêche partiellement de détecter un quelconque signal entre zones chez les glossines du Mouhoun. Le fait que la zone soit perturbée et que donc l'isolement puisse être récent entre les différentes zones peut également contribuer à brouiller l'image. En effet, alors que l'isolement par la distance est un phénomène qui se met très vite en place et devient détectable en quelques générations, comme le montrent nos simulations (Bouyer *et al.*, 2009), le F_{ST} met un certain nombre de générations à atteindre l'équilibre migration, mutation, dérive. Reprenons ces simulations.

Comprendre le manque de structure inter-zones avec un peu de simulations

Pour effectuer ces simulations, nous allons utiliser Easypop v 2.0.1 (Balloux, 2006, mise à jour de Balloux, 2001). Le problème avec les simulations, c'est de choisir un jeu de paramètres pertinent, car il y a une infinité de combinaisons possibles. Pour limiter notre travail, nous allons utiliser encore une fois les résultats de l'article de Rousset (1997) où l'on peut lire que dans un dispositif en une dimension on peut démontrer que, si N est le nombre d'individus d'un dème, m la proportion de migrants, D_e la densité efficace d'individus par km², σ la dispersion efficace (distance entre individus reproducteurs et leurs parents) et ε la distance entre deux dèmes adjacents :

$$Nm\varepsilon = D_e\sigma^2 \qquad (67)$$

La distance entre deux dèmes sera :

$$\varepsilon = \frac{D_e\sigma^2}{N_e m} \qquad (68)$$

En explorant les possibles (qui collent le mieux aux données), on peut obtenir $D_e\sigma^2 = 700$, $N = 30$ et $m = 0{,}5$ et donc $\varepsilon = 50$. Ce qui voudrait dire que la distance entre deux dèmes serait de l'ordre de 50 m. On va donc supposer que nos estimations étaient les meilleures dans la zone A et que nous y avons sous-estimé la taille des sous-populations. La distance entre les zones A et H est d'environ 70 km. Ces zones font environ 3 km chacune et nous allons simuler deux zones de même nature de 3 000 m, soit 3 000/50 ($\varepsilon = 50$, distance entre deux dèmes), 60 sous-populations chacune environ, séparées de 70 km, soit 1 400 sous-populations environ. Nous avons donc besoin de simuler 1 520 populations de taille 30 et échangeant 0,5 proportion de migrants dans un « stepping-stone » en une dimension.

Il faut maintenant lancer la simulation. Copiez Easypop dans le répertoire où vous souhaitez travailler, et double-cliquez dessus (sur le fichier programme pas sur le répertoire). Il faut ensuite répondre à toutes les questions. Vous souhaitez simuler des diploïdes à sexes séparés (dioïques) qui se croisent au hasard (on ne va pas se compliquer la vie). Nous voulons 1 520 populations de 30 individus avec un sexe-ratio équilibré (je dis bien UN sexe-ratio, car sexe est masculin en français et ratio de même en latin, quoiqu'en disent de nombreux écologues mal instruits), soit 15 femelles et 15 mâles. Vous souhaitez simuler un « stepping-stone » en une dimension tout au long de la simulation avec un taux de migration de 0,5 pour les deux sexes. Vous allez simuler 10 (plus rond que 7) loci indépendants avec un modèle de mutation KAM, 99 allèles possibles et un taux de mutation (le même pour tous) de 0,0001 qui correspond à un taux raisonnable, mais vous pourrez essayer avec 10^{-3} (consultez Ellegren, 2000 ; Balloux et Lugon-Moulin, 2002 ; Ellegren, 2004). Nous allons commencer avec une variabilité maximale (99 allèles équitablement répartis dans les 100 sous-populations), car cela fait gagner du temps (démarrer avec un seul allèle requiert un nombre important de générations avant d'obtenir quelque chose d'utilisable). Nous allons simuler 1 000 générations et récupérer tous les individus des 1 520 sous-populations. Nous ne voudrons pas connaître l'ascendance de nos individus. Nommez les fichiers résultats comme bon vous semble et ne demandez qu'une réplication (cela suffira ici). N'oubliez pas de valider chacun de vos choix par un retour chariot, sinon vous risquez d'attendre longtemps. En fonction de la puissance de votre ordinateur, la simulation durera plus ou moins longtemps (11 minutes avec ma double CPU 2.2 GHz avec 3.5 Go de RAM). Quand la simulation est terminée, Easypop vous demande le nom du fichier de sauvegarde des paramètres de la simulation (très utile !) et crée trois fichiers résultats : le détail de l'évolution de la simulation au cours des générations au niveau de divers paramètres (nombre d'allèles, diversité génétique, *F*-statistiques de Wright, etc.) (*.equ), un fichier de données Fstat (*.dat) et un fichier au format Genepop (*.gen). Il va falloir créer des fichiers pour tester l'effet « sous-structuration », l'effet Wahlund et l'interaction entre les deux. Pour l'effet sous-structuration, on ouvre le fichier .dat avec un bon éditeur de texte et on ne garde que six sous-populations dans les deux zones extrêmes « A-like », c'est-à-dire qu'on ne garde que les sous-populations 5-15-25-35-45-55 et 1 465-1 475-1 485 -1 495-1 505-1 515 que l'on recode de 1 à 12 en gardant bien à l'esprit que 1-6 = A1 et 7-12 = A2 (les deux zones extrêmes). Les populations marginales sont à éviter[13], c'est pourquoi on exclut les sous-populations 1 et 1 520. On crée un fichier de type HierFstat avec en première colonne la zone (1 et 2), en

[13] Dans un modèle en « *stepping-stone* » ouvert, les populations marginales ne reçoivent des migrants que des sous-populations qu'elles touchent (une seule en une dimension), alors que les sous-populations centrales reçoivent des migrants de tous les côtés.

deuxième la sous-population (1 à 12) et en colonnes 3 à 12 les génotypes aux 10 loci. On lance R et on charge le package hierfstat et on se met dans le répertoire où on a créé ce fichier que j'ai personnellement appelé MouhounA-001HierFstat.txt. Puis après les commandes habituelles :

```
> data<-read.table("MouhounA-001HierFstat.txt", header=TRUE)
> attach(data)
> loci<-data.frame(loc1,loc2,loc3,loc4,loc5,loc6,loc7,loc8,loc9,loc10)
> levels<-data.frame(Zone,Souspop)
> varcomp.glob(levels,loci)
```

on obtient le résultat suivant :

```
                   Zone     Souspop         Ind
Total         0.1296429   0.3276309   0.32083790
Zone          0.0000000   0.2274790   0.21967421
Souspop       0.0000000   0.0000000  -0.01010303
```

Nous pouvons remarquer que le F_{IS} est négatif, ce qui est normal puisque nous avons des sexes séparés. Ensuite, on a un fort F_{SZ} et un F_{ZT} beaucoup moins fort, comme prévu, mais cependant beaucoup plus important que celui observé entre zones pour nos données réelles de tsé-tsé. Nous allons refaire un fichier HierFstat en prenant deux zones adjacentes : sous-populations 5-15-25-35-45-55 (Zone 1) et 65-75-85-95-105-115 (Zone 2).

```
                   Zone     Souspop         Ind
Total        0.08133683   0.2849453   0.27293947
Zone         0.00000000   0.2216356   0.20856680
Souspop      0.00000000   0.0000000  -0.01679012
```

On diminue certes de beaucoup le F_{ZT}, mais pas autant que celui avec les sous-groupes définis par BAPS. Les clusters BAPS ne sont donc sans doute pas très au point. Il reste d'ailleurs un $F_{IS} > 0$ dans les clusters au lieu d'un excès. Il se peut donc que d'une part un effet Wahlund, combiné à des allèles nuls ainsi qu'à un isolement encore trop récent entre zones, expliquent le manque de différenciation observé entre zones, alors que l'on sait que les passages de mouches d'une zone à l'autre sont quasi impossibles (en principe). En fait, une analyse HierFstat en prenant les pièges comme sous-populations donne :

```
                   Zone        Trap          Ind
Total        0.01689790  0.07207679   0.2306512
Zone         0.00000000  0.05612733   0.2174274
Trap         0.00000000  0.00000000   0.1708918
```

La différenciation entre sous-échantillons (pièges) est fortement diminuée par rapport à l'attendu selon le modèle d'isolement par la distance, la différenciation entre

zones semble plus substantielle alors que la corrélation intra-individuelle (F_{IS}) est forte en raison d'un effet Wahlund et des allèles nuls et dans une moindre mesure du codage homozygote des mâles (peu nombreux en Zone A) aux loci hétérosomaux (dans les pièges le vrai F_{IS} = 0,14, voir plus haut).

CONCLUSIONS

Il y a une forte micro-structuration que le maillage des pièges ne permet pas de rendre avec précision, en particulier en ce qui concerne la taille des dèmes et leur distance entre eux, à moins que le voisinage soit strictement continu. Il n'en reste pas moins qu'un isolement par la distance a pu être mis en évidence et que l'on sait que la distance entre deux pièges doit être réduite, si l'on souhaite affiner notre vision de la structure des populations de *G. palpalis gambiensis* le long du Mouhoun (soit moins que 71 m entre pièges). Les estimations de densités et de dispersions convergent avec celles des données MRR, ce qui incite à une certaine confiance malgré les effets Wahlund, les petits échantillons et les allèles nuls qui ont très certainement considérablement brouillé les signaux. À l'avenir, pour les tsé-tsé de forêt galerie (le Mouhoun fait actuellement l'objet d'une campagne d'éradication par le PATTEC et ne sera donc pas étudiable avant longtemps), des pièges distants de 20 m, et le génotypage de davantage d'individus par piège sur des loci de meilleure qualité devraient permettre des estimations beaucoup plus précises encore. En particulier, la différenciation entre zones qui est apparue très incertaine pourra ainsi davantage être précisée et, par conséquent, une probabilité de dispersion entre ces zones estimée plus clairement.

6 Invasion de la Nouvelle-Calédonie par la tique du bétail *Rhipicephalus microplus* : hétérogénéité locale, dispersion et goulots d'étranglement

INTRODUCTION

Le jeu de données que nous allons analyser maintenant fait partie d'un projet finalisé en 2010 et publié dans quatre articles (Koffi *et al.*, 2006a ; Koffi *et al.*, 2006b ; Chevillon *et al.*, 2007a, b ; De Meeûs *et al.*, 2010). Nous allons refaire une partie de ces analyses et en ajouter quelques-unes. Nous pouvons ajouter ici que l'extraction d'ADN s'est montrée extrêmement délicate chez cette espèce et que nous avons dû travailler en aveugle (en ignorant s'il y avait effectivement des molécules dans nos extraits) avant d'obtenir les profils (pics). J'ai jugé cette précision importante pour les collègues qui décideraient de se lancer dans le génotypage de cette espèce. Notons que l'espèce *Rhipicephalus microplus* était communément nommée *Boophilus microplus* avant d'être remise en synonymie avec son nom actuel (Murrel et Barker, 2003). Le fichier de données se nomme "BoophilusAdultsDataCattle.txt".

ÉTAT DES LIEUX

Rhipicephalus microplus est une tique dure originaire du Sud-Est asiatique (Indonésie, Malaisie, Inde) (Labruna *et al.*, 2009). Cette tique a colonisé les zones intertropicales du monde entier en suivant l'introduction des bovins d'élevage et est aujourd'hui devenue une peste majeure des élevages de races européennes dans les agro-écosystèmes tropicaux et subtropicaux (Frisch, 1999 ; Jongejan et Uilenberg, 2004). *Rhipicephalus microplus* est considérée comme la tique la plus importante du monde du point de vue économique (Guerrero *et al.*, 2006) en y étant responsable de pertes de production directes (perte de poids par spoliation sanguine, surinfections et transmissions de maladies), ainsi qu'indirectes de par l'utilisation massive d'acaricides comme moyen de lutte (Frisch, 1999 ; Chevillon *et al.*, 2007b). Ajoutons que les coûts indirects se voient aggravés par l'évolution récurrente et extrêmement rapide de résistance aux différentes molécules utilisées dans les différentes parties de son aire de distribution (Frisch, 1999 ; Chevillon *et al.*, 2007b). Cette tique est dite monophasique, c'est-à-dire qu'elle accomplit son cycle de mues sur un seul individu hôte (en théorie). La femelle gravide, une fois son repas sanguin achevé, tombe au sol et meurt en libérant quelques milliers d'œufs sur le sol (environ 3 000) (Gallardo et Morales, 1999). Les larves qui éclosent attendent un hôte (un bovin

préférentiellement) pour s'y fixer et effectuer un premier repas sanguin, pour ensuite muer en nymphe sans quitter l'individu hôte et muer encore, après un second repas sanguin, en adulte. Le gardiennage pré-copulatoire peut s'établir dès que les femelles sont au stade nymphe (Falk-Vairant *et al.*, 1994), mais semble inefficace à empêcher les paternités multiples (Cutullé *et al.*, 2010). Après fécondation, les femelles se gorgent et se détachent ensuite de l'hôte pour tomber au sol et y mourir en y laissant leurs œufs. Bien que plutôt spécifique du genre *Bos* (Osterkamp *et al.*, 1999), on retrouve également *R. microplus* sur quelques autres hôtes (surtout Bovidae) (Hoogstraal et Aeschlimann, 1982), ainsi que sur le cheval (Ueti *et al.*, 2008) et sur le cerf rusa en Nouvelle-Calédonie (De Meeûs *et al.*, 2010).

En Nouvelle-Calédonie, l'espèce *R. microplus* a été introduite à partir de quelques individus en provenance d'Australie en 1942 (Verges, 1944 ; Rageau et Vergent, 1959), à la suite de quoi une quarantaine stricte a été instaurée. L'absence de l'espèce avant cette date et l'unicité de l'introduction est bien documentée (Bennett, 2004) et son origine australienne confirmée par analyse phylogénétique de l'ADN mitochondrial (Labruna *et al.*, 2009). *Rhipicephalus microplus* a ensuite rapidement colonisé tous les élevages de *Bos taurus* de l'île et est devenue résistante à tous les acaricides utilisés contre elle depuis (Ducornez *et al.*, 2005 ; Chevillon *et al.*, 2007b). Elle semble aussi commencer à s'adapter à un nouvel hôte, le cerf rusa (Barré *et al.*, 2001 ; De Meeûs *et al.*, 2010), lui même envahissant en Nouvelle-Calédonie, mais ceci est une autre histoire que nous n'aborderons pas ici.

Plusieurs questions se posent que la génétique des populations peut aborder sous un certain angle. Nous avons pour ce faire réalisé un échantillonnage, datant de 2003, de 698 tiques adultes prélevées sur vaches et génotypées au niveau de six loci (deux loci ont été rejetés, car donnant des résultats aberrants ; voir Koffi *et al.*, 2006a) dans huit élevages répartis sur l'île (fig. 62).

Le cycle spécial de cette tique permet de prédire une forte consanguinité à l'intérieur des individus hôtes due à la colonisation massive par une ou plusieurs pontes de tiques (frères et sœurs de la même ponte hautement apparentées entre elles et hétérogènes entre pontes différentes). On s'attend donc à une forte homozygotie relative moyenne intra-hôte (fort F_{IS}), mais très variable d'un hôte à l'autre (en fonction du nombre de fratries présentes) corrélée à une forte hétérogénéité inter-hôte dans chaque élevage (fort F_{SE}, avec S pour sous-population et E pour élevage). On comprend qu'ici, c'est l'individu hôte qui caractérisera la sous-population de tique ou, pour suivre la terminologie parasitologique, l'infra-population de tiques au sein de laquelle nous attendons une consanguinité importante.

Cette forte consanguinité conduit-elle à un biais de dispersion spécifique au sexe (Prugnolle et De Meeûs, 2002) ou/et à un évitement des conjoints apparentés ?

La diffusion apparemment rapide de la résistance (mais voir Chevillon *et al.*, 2007b) peut-elle être expliquée par les capacités dispersives de cette tique ?

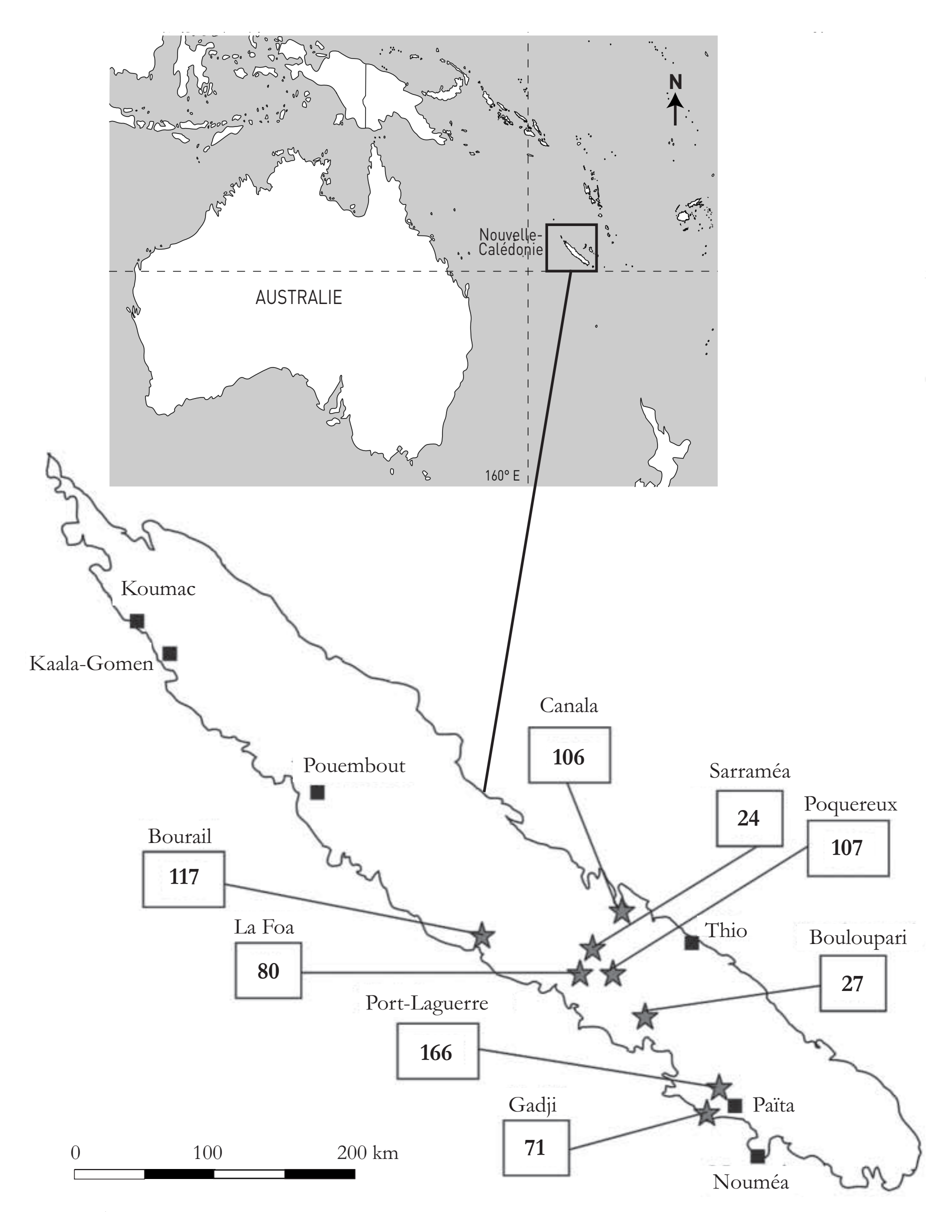


Figure 62
Sites et nombres de *Rhipicephalus microplus* adultes échantillonnées sur bétail en Nouvelle-Calédonie en 2003 et génotypées aux six marqueurs microsatellites.

Le goulot d'étranglement qu'a subi cette population lors de son introduction en 1942 est-il détectable à l'aide des marqueurs microsatellites mis au point par KOFFI *et al.* (2006b) ? Si oui, c'est que la quarantaine s'est montrée efficace, car des introductions multiples effacent la signature d'un goulot d'étranglement (CORNUET et

Luikart, 1996). Dans ce cas, en prenant quatre générations par an (Koffi *et al.*, 2006a), nous pouvons en déduire que ce goulot d'étranglement eut lieu il y a (2003-1942)×4 = 244 générations de tiques. Avec si peu de loci et des tailles d'échantillons de 30-50 individus environ, la possibilité de détecter un goulot d'étranglement n'est possible que si le paramètre τ de Cornuet et Luikart est compris entre 0,1 et 2,5 (Cornuet et Luikart, 1996 ; De Meeûs *et al.*, 2007a). Sachant que $\tau = t/2N_{eb}$, où t est le nombre de générations et N_{eb} est l'effectif efficace post-goulot d'étranglement, on peut en déduire, en cas de détection effective d'un goulot d'étranglement, que $N_{eb} = t/2\tau$ = [244/5, 244/0,2] = [49, 1220]. Cette gamme d'effectif efficace converge-t-elle avec les effectifs efficaces calculés à l'aide d'autres méthodes (Bartley *et al.*, 1992 ; Vitalis et Couvet, 2001a, b, c ; Balloux, 2004 ; Waples, 2006) ?

Et bien, c'est ce que nous allons rechercher ensemble.

ANALYSE DE LA CONSANGUINITÉ RELATIVE INTRA-HÔTE

Ce que nous allons rechercher ici, c'est la part prise par les infra-populations[14] de *R. microplus* dans la répartition de l'information génétique. Le paramètre à mesurer et à tester est donc le F_{SE} ou probabilité de fixation (homozygotie) dans les sous-populations (infra-populations) relative à celle des élevages. Nous aurons donc aussi besoin de mesurer le F_{IS}. C'est ce que nous ferons en premier, suivi des tests de déséquilibre de liaison, pour se débarrasser de ces derniers. Comme certains ont déjà dû le remarquer, j'ai en effet pris l'habitude de regarder d'abord ce qui se passe le plus localement avant d'essayer de comprendre ce qui se passe à des échelles plus globales.

Homozygotie et déséquilibre de liaison intra-hôte

Nous allons donc éditer "BoophilusAdultsDataCattle.txt" et ne garder que la colonne correspondant aux fermes et aux individus hôtes et celles des loci. Il faut recoder le label des individus hôtes en les fusionnant avec celui des fermes. N'oubliez pas que Fstat, que nous allons utiliser, n'accepte pas beaucoup de caractères pour les labels de population ou de loci. Notez que ferme et localité sont synonymes ici. Il faut recoder les allèles de chaque locus en les séparant pour pouvoir convertir ce fichier avec CREATE. En ce qui me concerne, cela donne le fichier que j'ai appelé "BoophilusAdultsDataCattleIndivHostFisLD.txt" (fig. 63).

[14] Voir la définition dans le glossaire.

Loca-Host»	B12»	B12»	C07»	C(
Boul-1»	192»	194»	145»	19
Boul-2»	200»	200»	179»	17
Boul-2»	192»	194»	179»	19
Boul-2»	192»	198»	145»	19
Boul-2»	192»	194»	179»	17
Boul-3»	198»	200»	145»	17
Boul-3»	194»	198»	179»	17
Boul-3»	194»	198»	145»	19
Boul-3»	194»	198»	145»	14
Boul-3»	194»	198»	179»	17
Boul-3»	194»	198»	145»	17
Boul-3»	198»	200»	145»	19
Boul-3»	194»	198»	179»	19
Boul-3»	194»	198»	145»	17
Boul-3»	192»	198»	145»	19
Boul-3»	194»	198»	179»	17
Boul-3»	198»	198»	145»	14
Boul-3»	194»	198»	150»	19
Boul-3»	194»	198»	145»	15
Boul-3»	194»	194»	145»	14
Boul-3»	194»	198»	179»	19
Boul-3»	194»	194»	145»	19
Boul-3»	194»	198»	145»	19
Boul-3»	194»	198»	0»	0»
Boul-3»	194»	198»	145»	17
Boul-3»	194»	198»	150»	15
Boul-3»	194»	200»	179»	17
Bour-1»	198»	198»	179»	17
Bour-1»	194»	198»	179»	19
Bour-1»	194»	200»	150»	17
Bour-1»	194»	194»	145»	17
Bour-1»	192»	196»	0»	0»
Bour-1»	194»	198»	145»	14
Bour-1»	194»	194»	0»	0»
Bour-1»	194»	198»	150»	18
Bour-1»	198»	198»	150»	17
Bour-1»	198»	198»	179»	17
Bour-1»	194»	198»	145»	14
Bour-1»	192»	194»	145»	17
Bour-1»	198»	198»	145»	19
Bour-1»	194»	198»	145»	14
Bour-1»	192»	198»	145»	17

Figure 63
Extrait du fichier de données pour l'analyse F_{IS} et déséquilibre de liaison de *R. microplus* de bétail en Nouvelle-Calédonie (>> signale une tabulation).

Nous allons convertir ce fichier au format Fstat par l'entremise de CREATE (comme au chapitre précédent). Une fois cela fait, et avant de lancer Fstat, il faut éditer le fichier "BoophilusAdultsDataCattleIndivHostFisLD-FSTAT-POPULATION NAMES.lab" (nom des sous-populations) que vient de créer CREATE pour supprimer les deux dernières colonnes (je ne sais pas pourquoi CREATE fait ça). Profitons-en pour raccourcir le nom des fichiers en "BoophilusAdultsDataCattleIndivHostFisLD.dat" et "BoophilusAdultsDataCattleIndivHostFisLD.lab". On lance Fstat, on charge

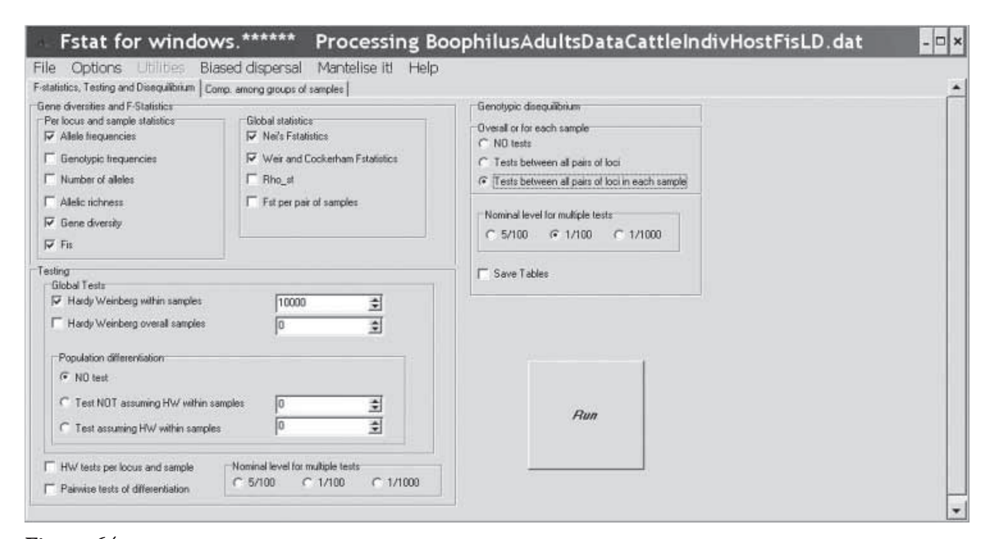


Figure 64
Cases à cocher dans Fstat pour l'analyse F_{IS} et déséquilibre de liaison par paire de loci des données microsatellites de *R. microplus*.

"BoophilusAdultsDataCattleIndivHostFisLD.dat" et le fichier associé ".lab". On coche les options comme indiqué dans la figure 64.

Ensuite, on clique sur "Run" et on attend la fin des permutations (7 mn sur ma machine, vous avez le temps de consulter vos courriels ou de faire une partie de démineur).

Dans le fichier ".out", nous remarquons qu'aucun locus n'a un allèle de fréquence trop dominante (pas de fréquence moyenne supérieure à 0,5 dans le cas présent). Les tests de déséquilibre de liaison sont donc « raisonnablement puissants ». Un seul de ces tests sur les 15 paires de loci possibles est significatif au seuil 5 %. Ceci n'est pas significativement différent de l'attendu sous l'hypothèse nulle, comme nous le donne l'utilisation de MultiTest avec alpha = 0,05, k = 15 et k' = 1 (*P*-value = 0,537) ou la commande R "binom.test(1, 15, 0,05, alternative="greater")". Il n'y a donc pas de signal significatif de déséquilibre de liaison à ce niveau. Nous pouvons considérer ces loci comme statistiquement indépendants.

Les résultats de l'analyse des F_{IS} sont représentés dans la figure 65.

Normalement, à ce stade, ces analyses ne doivent vous poser aucun problème. Nous constatons qu'un déficit en hétérozygotes très significatif, quoique léger (F_{IS} = 0,04), existe au sein des infra-populations de *R. microplus* en Nouvelle-Calédonie. Une légère variation de ce F_{IS} entre loci (fig. 65) pourrait suggérer l'influence d'allèles nuls dans ce déficit sauf que les loci responsables de cette variation (D12 et D10) ne semblent pas influencer beaucoup le résultat global. Certaines infra-populations étant de petites tailles, il paraît délicat d'utiliser MicroChecker à ce stade. Mais la

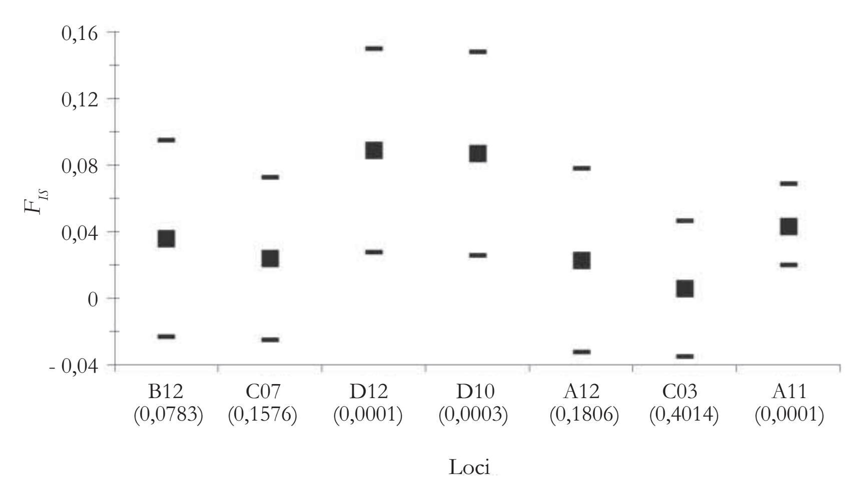


Figure 65
Résultat des analyses d'homozygoties relatives des individus (F_{IS}) au niveau des infra-populations (dans chaque individu hôte) de *R. microplus* sur bovins en Nouvelle-Calédonie. Les intervalles de confiance (95 %) sont issus de jackknives sur populations où le F_{IS} est calculable (33 infra-populations) avec la méthode décrite en p. 72-73 (1re partie), sauf pour la valeur moyenne (All) dont l'intervalle de confiance correspond à 5 000 bootstraps effectués par Fstat. La probabilité de dévier de 0 sous H_0 (*P*-value obtenue après 10 000 permutations des allèles entre individus de la même infra-population) est donnée entre parenthèses.

méthode de régression des F_{IS} en fonction du nombre de blancs par locus et sous-population, que nous avons déjà utilisée en p. 197-199 de cette partie, peut être aisément réalisée. Ici, nul test n'est nécessaire étant donné que la corrélation est très faible et de toutes manières négative entre les deux variables. Les allèles nuls ne semblent pas pouvoir expliquer nos résultats. Nous allons laisser cela de côté et vérifier tout d'abord que le niveau infra-population est pertinent.

Analyse hiérarchique

Nous allons donc avoir besoin de HierFstat et de coder un fichier en ce sens, avec comme facteurs (du moins inclusif au plus inclusif) : la Nouvelle-Calédonie (T), l'élevage E), l'infra-population (S) et l'individu tique (I). Reprenons donc le fichier “BoophilusAdultsDataCattle.txt” et recodons-le afin d'obtenir quelque chose de la forme (fig. 66).

Remarquez que j'ai recodé les génotypes avec deux chiffres au lieu de trois, car sinon HierFstat me retournait un message d'erreur (mais je ne sais pas pourquoi, car normalement ça doit marcher avec trois chiffres par allèle). N'oubliez surtout pas de recoder les données manquantes “0000” en “NA”. Il faut ensuite ouvrir R, on charge

Farm	Host	B12	C07	D12
1	1	0102	0108	0209
1	2	0505	0606	0709
1	2	0102	0608	0209
1	2	0104	0108	0209
1	2	0102	0606	0204
1	3	0405	0106	0912
1	3	0204	0606	0709
1	3	0204	0108	0204
1	3	0204	0101	0205
1	3	0204	0606	0909
1	3	0204	0106	0709
1	3	0405	0108	0202
1	3	0204	0608	0209
1	3	0204	0105	0209
1	3	0104	0108	0204
1	3	0204	0606	0709
1	3	0404	0101	0205
1	3	0204	0208	0609
1	3	0204	0102	0909
1	3	0202	0101	0209
1	3	0204	0608	0202
1	3	0202	0108	0209
1	3	0204	0108	0209
1	3	0204	NA	0209
1	3	0204	0106	0209
1	3	0204	0202	0909
1	3	0205	0606	0505
2	4	0404	0606	0909
2	4	0204	0608	0505
2	4	0205	0206	0205
2	4	0202	0106	0204
2	4	0103	NA	0204
2	4	0204	0101	0202
2	4	0202	NA	0202
2	4	0204	0207	0202
2	4	0404	0206	0202
2	4	0404	0606	0202
2	4	0204	0101	0409

Figure 66
Extrait du fichier "BoophilusAdultsDataCattleHierFstat.txt" pour l'analyse des *F* hiérarchiques par HierFstat.

le "package hierfstat", on se met dans le bon répertoire et on tape les commandes habituelles :

```
data<-read.table("BoophilusAdultsDataCattleHierFstat.txt", header=TRUE)
attach(data)
loci<-data.frame(B12,C07,D12,D10,A12,C03)
> levels<-data.frame(Farm,Host)
> varcomp.glob(levels,loci)
```

ce qui donne le résultat :

```
                    Farm         Host         Ind
Total         0.01535231  0.016482637  0.05917112
Farm          0.00000000  0.001147949  0.04450201
Host          0.00000000  0.000000000  0.04340389
```

que l'on teste avec les commandes :

```
> test.within(loci,test=Host,within=Farm,nperm=1000)
```

ce qui renvoie à une *P*-value = 0,132 et

```
> test.between(loci,rand.unit=Host,test=Farm,nperm=1000)
```

ce qui renvoie une *P*-value = 0,001

Il en ressort que si le niveau hôte (infra-population de parasites) explique une part infime et non significative de la variation génétique, le niveau ferme est quant à lui très significatif. Nous allons donc recommencer en ignorant le niveau infra-population.

ANALYSES INTRA ET INTER-FERME

Homozygotie, déséquilibre de liaison intra-ferme et différentiation globale

Vous allez donc recréer un fichier Fstat, mais uniquement avec les fermes (localités). Ce fichier, BoophilusAdultsDataCattleFarmFisLD.dat, nous allons l'analyser comme indiqué en figure 67. Il en ressort qu'aucun test de déséquilibre de liaison n'est significatif (*P*-value > 0,079), ce qui confirme que le résultat avec les infra-populations n'était pas dû à un manque de puissance causé par les faibles tailles de ces infra-populations. Le F_{IS} est toujours très significativement (*P*-value = 0,0001) au dessus de 0 à F_{IS} = 0,044, soit sensiblement la même valeur qu'avant, ce qui

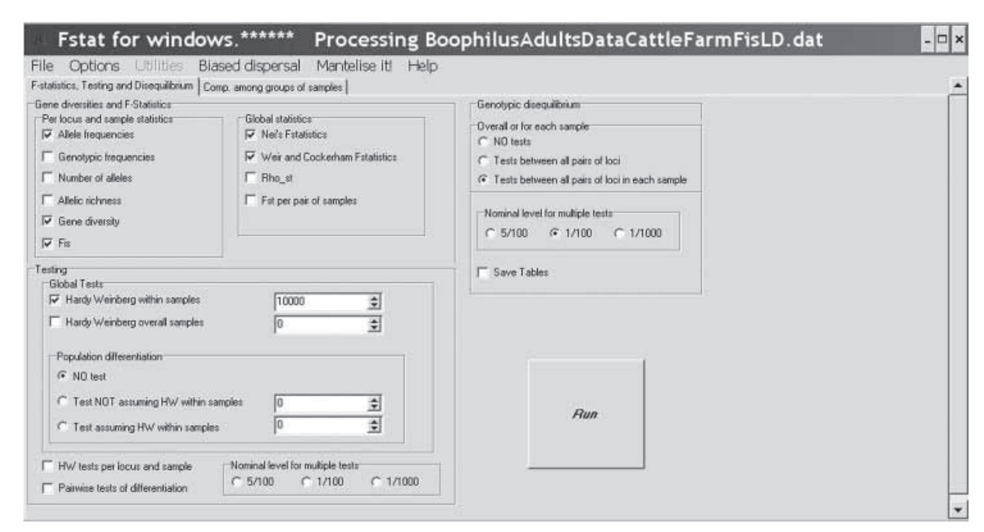


Figure 67
Cadre de Fstat avec les analyses à effectuer
pour les données des marqueurs microsatellites de *Boophilus microplus.*

confirme que réunir les infra-populations d'une même ferme est valide (pas d'effet Wahlund). Enfin, la différenciation entre fermes est significativement supérieure à 0 (P-value = 0,0001) avec un F_{ST} = 0,016, ce qui, compte tenu de la diversité génétique présente H_s = 0,704, représente une différentiation standardisée relativement faible de $F_{ST}' = F_{ST}/(1 - H_s)$ = 0,05 et suggère une importante migration entre fermes.

Analyse des biais de dispersion sexe-spécifiques

Trois types d'analyses sont possibles ici. Soit une analyse par élevage pour tester le biais de dispersion spécifique au sexe entre infra-populations (huit analyses), soit une analyse sur l'ensemble des infra-populations où il y a des mâles et des femelles (33 infra-populations en tout), soit une analyse sur l'ensemble des fermes sans distinguer les infra-populations, soit donc 10 analyses Fstat en tout. Il faut repartir du fichier source pour recoder les données au format requis (fig. 68). Notez que les allèles doivent être codés avec deux chiffres pour ces analyses.

```
2»                  6»                  12¶
B12,·C07,·D12,·D10,·A12,·C03¶
pop¶
F,Boul-2,»          0505»               0606»               0709»
F,Boul-2,»          0102»               0608»               0209»
M,Boul-2,»          0104»               0108»               0209»
M,Boul-2,»          0102»               0606»               0204»
pop¶
F,Boul-3,»          0405»               0106»               0912»
F,Boul-3,»          0204»               0606»               0709»
F,Boul-3,»          0204»               0108»               0204»
```

Figure 68
Exemple d'un fichier pour l'analyse de biais de dispersion sexe-spécifique entre infra-populations de *B. microplus* dans l'élevage de Bouloupari.

Vous lancez Fstat et vous choisissez le menu déroulant “Biased dispersal”. Cochez les paramètres “Mean assignment” (AI_c), “Variance of assignment” (vAI_c) et “Fst” (F_{ST}) qui sont les plus performants, comme discuté ailleurs (p. 91-92 de la première partie et p. 147-153 de la seconde partie). Les tests doivent être bilatéraux et on procédera à 10 000 permutations. Les résultats des analyses par élevage sont présentés dans le tableau 30.

On y voit bien qu'aucun signal n'existe. Il n'y a que deux tests significatifs sur les 24 effectués, ce qui n'est pas significativement différent des 5 % attendus sous l'hypothèse nulle (test binomial, P-value = 0,34). De plus, il y a de nettes contradictions entre paramètres pour un même site ou entre sites pour un même paramètre. Les analyses sur l'ensemble des infra-populations ou sur l'ensemble des fermes en ignorant les infra-populations confirment l'absence de tout signal (P-value > 0,27). Il n'y a donc aucune trace d'un biais de dispersion spécifique au sexe chez cette tique.

Tableau 30
Résultats des analyses de biais de dispersion spécifique au sexe entre infra-populations de *B. microplus* au sein des élevages de Nouvelle-Calédonie. Les valeurs de paramètres donnant le sexe (F ou M) le moins dispersant sont en gras et les *P*-values ≤ 0,05 sont en italique.

		AI_c	vAI_c	F_{ST}
Bouloupari	F	- 0,17979	**2,91388**	**0,0384**
	M	**0,20975**	3,63331	0,0227
	P-Value	0,6011	0,6718	0,8139
Bourail	F	- 0,02319	**3,97564**	- 0,0081
	M	**0,02203**	4,44469	**- 0,0078**
	P-Value	0,9042	0,77	0,9827
Canala	F	- 0,10075	3,67074	- 0,0158
	M	**0,10075**	**3,48**	**0,0099**
	P-Value	0,5944	0,8258	0,0553
Gadji	F	**0,08235**	**4,02626**	**0,0068**
	M	- 0,1342	5,1342	- 0,0229
	P-Value	0,6989	0,6833	0,191
La Foa	F	- 0,44351	**2,77627**	- 0,0037
	M	**0,42187**	3,04567	**0,0041**
	P-Value	*0,0349*	0,8009	0,6625
Poquereux	F	- 0,03302	4,10552	0,**0031**
	M	**0,04549**	**2,65887**	- 0,0065
	P-Value	0,8271	0,1071	0,5144
Port-Laguerre	F	**0,0633**	4,95572	- 0,0082
	M	- 0,06179	**4,53941**	**0,0026**
	P-Value	0,7281	0,7274	0,2289
Sarraméa	F	- 0,0785	1,26806	**0,1069**
	M	**0,0785**	**1,15348**	- 0,0136
	P-Value	0,7546	0,9723	*0,0103*

Tests de pangamie

Ces données ne sont disponibles que pour quatre sites et c'est pourquoi elles sont disponibles dans un fichier à part "BooCattleCouples.txt" dans lequel figure le nom du couple auquel appartiennent chaque femelle et chaque mâle. Il s'agit de tester si les couples s'associent de façon consanguine, ce qui pourrait expliquer le déficit en hétérozygotes observé. Rappelons que selon l'équation 66 (voir aussi réponse 11), le taux de croisement frères-sœurs nécessaire à expliquer un F_{IS} = 0,044 se déduit de :

$$b = \frac{4F_{IS}}{1+3F_{IS}} = 0{,}16$$

Pour explorer le rôle possible d'un appariement entre apparentés, nous allons tester s'il y a pangamie (appariement au hasard dans nos données). Nous allons utiliser pour ce faire la même technique que celle développée par Prugnolle *et al.*, 2004b. Il s'agit d'un test de Mantel de corrélation entre deux matrices : une matrice décrivant l'apparentement entre chaque paire d'individus de sexes différents et une matrice décrivant le statut apparié (1) ou non apparié (0) des individus. Étant donné qu'il y a une différenciation génétique substantielle entre fermes, nous devrons entreprendre quatre tests séparés (un par ferme). Cependant, comme certaines fermes possèdent énormément d'individus génotypés (Port-Laguerre) où le test de Mantel de Fstat ne marchera pas et par souci d'homogénéité, nous travaillerons par individu hôte (autant de tests que d'hôtes disponibles sur l'ensemble de l'échantillonnage). Nous allons mesurer l'apparentement entre individus avec le logiciel ML Relate (Kalinowski *et al.*, 2006) (téléchargeable librement à http://www.montana.edu/kalinowski/Software.htm). Ce logiciel prend directement des fichiers Genepop. Construisez autant de fichiers qu'il y a d'infra-populations comme dans l'exemple qui suit (fig. 69).

Ensuite, il faut lancer le logiciel ML-Relate, aller au menu déroulant "File" et charger votre fichier, ce qui fait apparaître les fréquences alléliques. Puis vous cliquez sur le

```
Couples·de·Canala·B1¶
B12¶
C07¶
D12¶
D10¶
A12¶
C03¶
POP¶
b1C09F»    ,»    9294»    4550»    9210»    5252»    9697»    4652¶
b1C09M»    ,»    9298»    7985»    1211»    5253»    9797»    4658¶
b1C11F»    ,»    9898»    7979»    1213»    5254»    9696»    4646¶
b1C11M»    ,»    9698»    4585»    9213»    5254»    9799»    5254¶
b1C12F»    ,»    9420»    4579»    9292»    5254»    9697»    5254¶
b1C12M»    ,»    9898»    6879»    1113»    5255»    9797»    4658¶
b1C14F»    ,»    9298»    4550»    9292»    5252»    9999»    4868¶
b1C14M»    ,»    9898»    4550»    1414»    5252»    9999»    4854¶
b1C18F»    ,»    9498»    7985»    1212»    5252»    9697»    5454¶
b1C18M»    ,»    9494»    4579»    9292»    5052»    9595»    4652¶
b1C19F»    ,»    9498»    4585»    9292»    5252»    9699»    5254¶
b1C19M»    ,»    9298»    4579»    9213»    5052»    9799»    5258
```

Figure 69
Exemple de données pour ML-RELATE
pour mesurer l'apparentement entre tiques adultes du premier bovin à Canala.

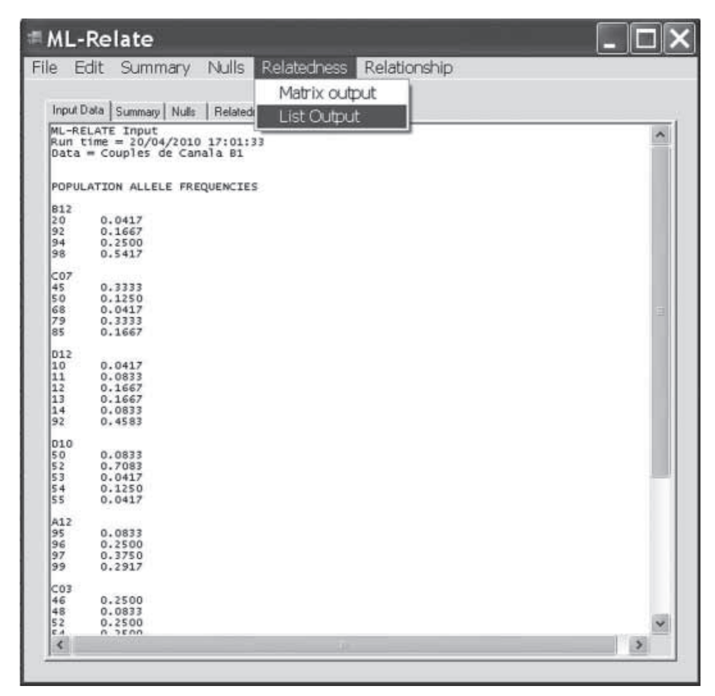


Figure 70
Menu ML-Relate à choisir.

menu déroulant “Relatedness” et choisissez “List Output” (fig. 70), car votre matrice ne sera pas carrée, il faudra donc présenter les données au format colonnes à Fstat. C’est un détail qui a son importance, la procédure de Fstat est issue de RT de Manly (Manly, 1997) qui permet d’effectuer des tests de Mantel entre matrices non carrées (impossible avec Genepop, par exemple), ce qui est bien commode.

Ceci vous donne toutes les paires d’apparentement que vous devez sélectionner avec la souris comme dans la figure 71. Copiez ces données et collez-les dans un logiciel qui vous permettra de trier ces données. Vous allez en effet devoir ne garder que les couples réalisés et potentiels. La première colonne ne contiendra donc que les femelles et la seconde que les mâles.

Votre fichier intermédiaire doit donc ressembler à la figure 72. On y voit bien que les données ont été triées par sexe pour le premier et le deuxième individu de la paire et que seules les femelles ont été gardées pour le premier et les mâles pour le second membre de chaque paire. Une dernière colonne a été créée pour donner le statut accouplé (1) ou non accouplé (0) de la paire. Ceci est facilement obtenu sous Excel par une formule conditionnelle “si(coordonnées case 1=coordonnées case 2; 1; 0)” (fig. 73).

Il faut ensuite mettre ce fichier au format acceptable pour le test de Mantel de Fstat. Ceci est très bien expliqué dans l’aide en ligne du logiciel et je ne m’y attarderai donc

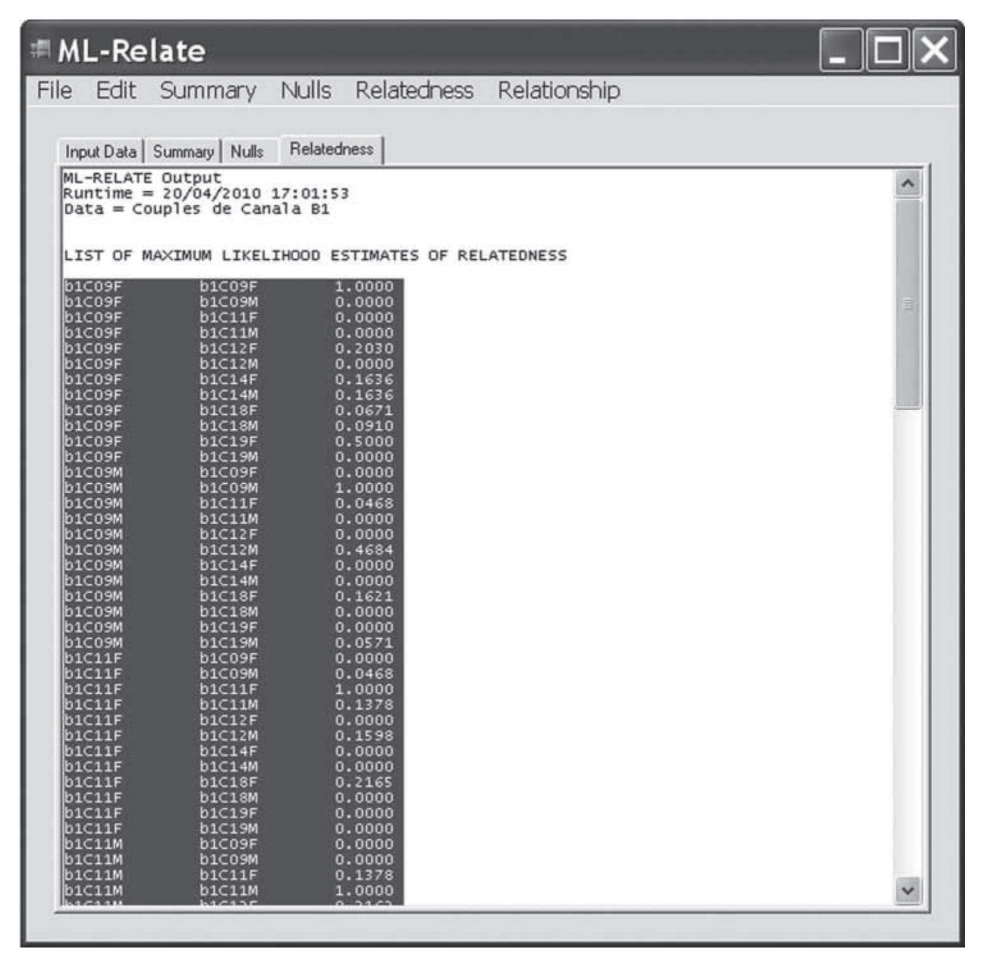


Figure 71
Sélection des résultats de calculs d'apparentement pour les *B. microplus* du premier bovin de Canala dans la fenêtre de ML-Relate.

pas. Lancez Fstat et allez directement dans le menu "Mantelize it" et dans le menu "File", chargez votre fichier. Le logiciel vous demande alors un fichier de sortie (résultats). Personnellement, je prends le même nom, mais je mets l'extension ".man". Une nouvelle fenêtre apparaît. Il vous faut choisir la variable dépendante qui est ici le statut du couple. Sélectionnez donc "Couple" et mettez-le dans la case "Dependant" avec le bouton ">" comme indiqué dans la figure 73. Apparaît alors la case de la variable explicative qu'il faut remplir avec "R" le coefficient d'apparentement. Tapez 10 000 pour le nombre de randomisations et sur "Run" comme dans la figure 74.

Dans les résultats, ne gardez ici que la valeur de corrélation (0,18 ici) et celle de la *P*-value (0,28). Constatez que cette *P*-value est une *P*-value bilatérale. Or nous

F2 =SI(A2=C2;1;0)

	A	B	C	D	E	F
1	Tick1	Sex1	Tick2	Sex2	R	Couple
2	b1C09	F	b1C09	M	0	1
3	b1C09	F	b1C11	M	0	0
4	b1C09	F	b1C12	M	0	0
5	b1C09	F	b1C14	M	0.1636	0
6	b1C09	F	b1C18	M	0.091	0
7	b1C09	F	b1C19	M	0	0
8	b1C11	F	b1C09	M	0.0468	0
9	b1C11	F	b1C11	M	0.1378	1
10	b1C11	F	b1C12	M	0.1598	0
11	b1C11	F	b1C14	M	0	0
12	b1C11	F	b1C18	M	0	0
13	b1C11	F	b1C19	M	0	0
14	b1C12	F	b1C09	M	0	0
15	b1C12	F	b1C11	M	0.2162	0
16	b1C12	F	b1C12	M	0	1
17	b1C12	F	b1C14	M	0	0
18	b1C12	F	b1C18	M	0.1775	0
19	b1C12	F	b1C19	M	0	0
20	b1C14	F	b1C09	M	0	0
21	b1C14	F	b1C11	M	0	0
22	b1C14	F	b1C12	M	0	0
23	b1C14	F	b1C14	M	0.4694	1
24	b1C14	F	b1C18	M	0	0
25	b1C14	F	b1C19	M	0.0429	0
26	b1C18	F	b1C09	M	0.1621	0
27	b1C18	F	b1C11	M	0	0
28	b1C18	F	b1C12	M	0	0
29	b1C18	F	b1C14	M	0	0
30	b1C18	F	b1C18	M	0	1
31	b1C18	F	b1C19	M	0	0
32	b1C19	F	b1C09	M	0	0
33	b1C19	F	b1C11	M	0.5556	0
34	b1C19	F	b1C12	M	0	0
35	b1C19	F	b1C14	M	0	0
36	b1C19	F	b1C18	M	0.0946	0
37	b1C19	F	b1C19	M	0	1

Figure 72
Aspect du fichier pour le test de Mantel de corrélation entre apparentement et accouplement chez *Rhipicephalus microplus* de la première vache de Canala.

recherchons un signal spécifique susceptible d'expliquer nos déficits en hétérozygotes et donc une corrélation positive ($R > 0$). Nous devrons donc transformer ces *P*-value en les divisant par deux pour celles dont le $R > 0$, ou en posant 1-(*P*-value/2) pour celles dont la corrélation est négative. Ce n'est pas idéal, mais ça doit coller à peu près.

Il faut recommencer avec chacune des infra-populations de tous les hôtes de tous les sites.

L'ensemble des résultats est synthétisé dans le tableau 31. Nous y voyons que le signal, même s'il est assez faible, est cependant significatif (*P*-value = 0,0466 en bilatéral,

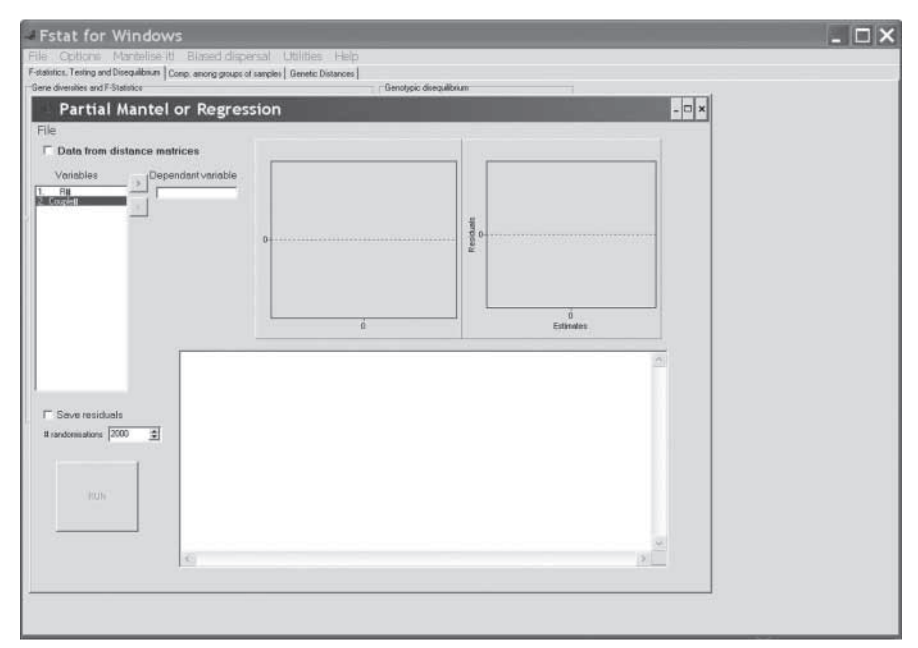


Figure 73
Sélection de la variable dépendante dans le menu “Mantelise it” de Fstat.

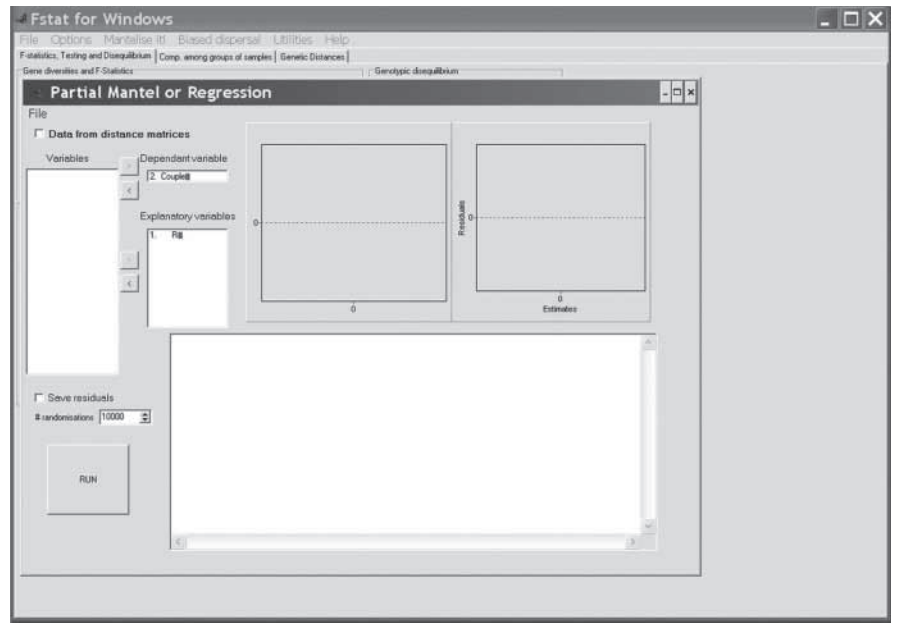


Figure 74
Seconde étape pour le Mantel avant de cliquer sur “Run”.

Tableau 31
Tableau des résultats des tests de corrélation (*R*) de Mantel entre l'apparentement et l'accouplement des tiques des infra-populations de *Rhipicephalus microplus*. Les tests au départ bilatéraux ont été unilatéralisés dans le sens *R* > 0 (sens recherché) en divisant la *P*-value unilatérale par deux et en la retranchant de 1 pour celles correspondant aux corrélations négatives. Pour le total, la corrélation est la moyenne non pondérée sur l'ensemble des 20 infra-populations et les *P*-values ont été combinées par la méthode binomiale généralisée de MultiTest (De Meeûs *et al.*, 2009) avec *k*' = *k*/2 = 10. La 10[e] *P*-value en ordre croissant est indiquée en gras.

Vache	*R*	*P*-value bilatérale	*P*-value unilatérale
Bourail, bovin 1	- 0,004306	0,9084	0,5458
Bourail, bovin 2	- 0,030853	0,4458	0,7771
Bourail, bovin 3	0,067008	0,1937	0,09685
Bourail, bovin 4	0,039708	0,3115	0,15575
Bourail, bovin 5	- 0,041236	0,322	0,839
Canala, bovin 1	0,181052	0,2836	0,1418
Canala, bovin 2	- 0,141193	0,1078	0,9461
Canala, bovin 4	0,237409	0,0057	0,00285
Canala, bovin 5	- 0,064224	0,4947	0,75265
Canala, bovin 6	0,087719	**0,2988**	0,1494
La Foa, bovin 1	0,195527	0,2515	0,12575
La Foa, bovin 2	- 0,006741	0,9561	0,52195
La Foa, bovin 3	0,04795	0,7655	0,38275
La Foa, bovin 4	0,070247	0,4543	0,22715
La Foa, bovin 5	0,273734	0,0884	0,0442
Port-Laguerre, bovin 1	- 0,033541	0,2091	0,89545
Port-Laguerre, bovin 2	- 0,003524	0,8774	0,5613
Port-Laguerre, bovin 3	0,062013	0,0243	0,01215
Port-Laguerre, bovin 4	- 0,041535	0,1094	0,9453
Port-Laguerre, bovin 5	0,016508	0,534	**0,267**
Total	0,0455861	0,0468	0,0221

P-value = 0,0219 en unilatéral). Si cette corrélation est due à des croisements entre frères et sœurs au taux (voir plus haut) de 16 %, on aurait obtenu une corrélation beaucoup plus forte. Par exemple, avec 20 couples réalisés dont 16 % (donc 3) ont un apparentement de 0,522, car des pleins frères de consanguinité F_{IS} ont un apparentement de $1/4\times(1 + F_{IS})\times2$, et le reste un apparentement de $\sim 2F_{IS} = 0{,}088$ (voir réponse 14), on obtient sur l'ensemble des 400 couples (possibles et réalisés) une corrélation de 0,35 et une *P*-value bilatérale de 0,0095. Il doit donc exister une autre explication pour rendre compte de l'entièreté du F_{IS} des populations de cette tique. Comme pour les tiques *I. ricinus*, il existe peut-être une structure cachée, un effet Wahlund. Nous allons donc, dans la section qui va suivre, rechercher cet effet.

Recherche d'un effet Wahlund

Nous allons ici de nouveau réutiliser le logiciel BAPS que nous ferons fonctionner dans chaque ferme étant donné que nous avons montré que les tiques se répartissent au hasard dans ces élevages, mais pas entre élevages. Le format et la procédure ayant déjà été décrits en détail, nous allons tout de suite regarder les résultats en termes de nombre de clusters trouvés et de leur F_{IS}. Nous allons aussi en profiter pour regarder les résultats obtenus avec un nouveau logiciel, Flock (Duchesne *et al.*, 2010 ; Duchesne et Turgeon, 2009), qui n'existait pas encore quand j'ai commencé à rédiger ce manuel (et oui j'ai mis un temps fou !).

BAPS

Les partitions obtenues par BAPS dans les différents sites donnent des résultats plus ou moins bons avec parfois beaucoup de clusters (tabl. 32). Le F_{IS} de cette nouvelle partition chute de façon spectaculaire ($F_{IS} = -0{,}14$ avec un intervalle de confiance à 95 % de - 0,17 à - 0,11, contre 0,04 compris entre 0,02 et 0,07 pour les données initiales). Une telle chute vers des valeurs aussi négatives est difficile à expliquer par un simple effet Wahlund. Cela signifierait en effet que chaque élevage renferme de nombreuses et minuscules sous-populations de tiques (de taille $N_e \sim 4$ selon Balloux, 2004 équation 12).

Une autre piste est celle de la présence de plusieurs individus de la même ponte (frères et sœurs) répartis sur l'ensemble des bovins d'une ferme. Cela peut arriver si la variance de survie entre pontes est très forte de telle sorte qu'à chaque génération ne restent dans un élevage donné que les représentants de quelques pontes, avec beaucoup de survivants par ponte. Cette hypothèse n'est pas incompatible avec le fait que les élevages subissent des traitements récurrents et est en accord avec le F_{IS} effectivement attendu très négatif dans ce cas (voir réponse 15). En appliquant le modèle de la réponse 15 aux données par locus et par élevage et en faisant la moyenne pondérée par locus on obtient en effet, pour des fratries, un F_{IS} compatible avec les résultats donnés par BAPS (fig. 75), mais significativement différent des données non manipulées.

Tableau 32
Nombre de clusters BAPS, effectifs par site et *P*-value donnée par BAPS (probabilité pour que la partition soit bonne) pour les différents sites. Les nombres de clusters obtenus par Flock et par le critère d'optimisation décrit dans la documentation (Flock optimisé, "K estimates based on plateau analysis" non discuté ici) sont aussi indiqués.

Élevage	BAPS	N	*P*-value	Flock	Flock optimisé
Bouloupari	5	27	0,43	4	2
Bourail	17	117	0,46	12	1
Canala	13	106	0,49	11	1
Gadji	11	71	0,69	8	1
La Foa	10	80	0,16	8	2
Poquereux	15	107	0,62	11	1
Port-Laguerre	20	166	0,40	15	1
Sarramea	8	24	0,28	4	1

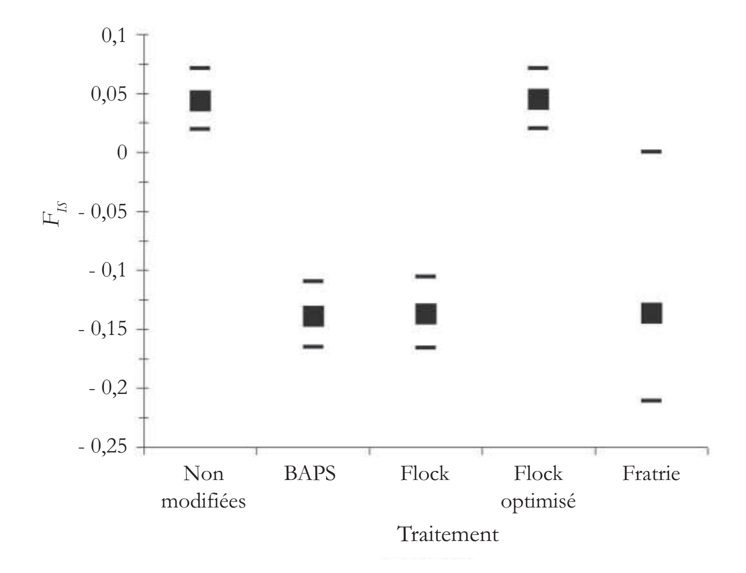


Figure 75
F_{IS} obtenus pour le jeu de données non modifiées (par ferme), pour le jeu de données clusterisées par BAPS, par Flock (nombre maximum de clusters) et Flock optimisé (K estimates based on plateau analysis), ainsi que pour l'attendu pour une structure en fratrie (modèle de la Réponse 15) avec les fréquences alléliques par ferme. Les intervalles de confiance (95 %) sont obtenus par bootstrap sur les loci sauf pour le F_{IS} des fratries obtenu avec la valeur maximale et minimale observées sur les moyennes (pondérées sur l'ensemble des fermes) par locus.

Flock

Je ne vais pas détailler ici l'analyse, car je manque de toutes manières de recul sur ce programme, mais je trouvais intéressant d'évoquer ici ce nouveau logiciel, qui n'est de toutes manières pas très difficile à utiliser. Les résultats donnés par Flock sont comparables à ceux obtenus par BAPS, mais avec moins de clusters (BAPS a en effet tendance à exagérer le nombre de clusters, LATCH *et al.*, 2006) pour ce qui est du nombre maximal de clusters obtenus (tabl. 32, fig. 75). Pour le minimum de clusters (Flock optimisé) par contre, les résultats ne donnent pas grand-chose d'exploitable.

CONCLUSION DES ANALYSES INTRA-FERMES

L'ensemble de nos résultats suggère une libre circulation des tiques entre hôtes de la même ferme, mais un isolement des fermes qu'il convient d'analyser plus en détail (voir plus bas). Cette libre circulation contredit le modèle classique de fidélité stricte des individus tiques vis-à-vis de l'individu hôte colonisé par les larves et explique bien comment, malgré une transmission transovarienne négligeable, *R. microplus* reste un vecteur majeur d'*Anaplasma marginale*, une bactérie très pathogène du bétail en zones intertropicales (UILENBERG, 1976) (pathogène absent de Nouvelle-Calédonie). Du stade larvaire aux adultes, des échanges de tiques ont donc probablement lieu entre individus hôtes, vraisemblablement lors de contacts physiques entre bêtes. Ce phénomène est couplé avec une structure en fratries combinée à des accouplements légèrement assortis génétiquement. Ceci provient possiblement du fait que les larves issues d'une même ponte ont plus de chances d'atteindre la maturité sexuelle en même temps ce qui, couplé avec une variance de survie importante, crée un léger, mais très significatif effet Wahlund.

ISOLEMENT PAR LA DISTANCE

Comme nous disposons des coordonnées GPS des sites, nous allons les utiliser dans le logiciel Genepop 4 (ROUSSET, 2008). Les données (fichier texte) doivent se présenter comme dans la figure 76. Genepop 4 doit être copié dans le répertoire de travail. Cliquez deux fois sur le logiciel. Une fenêtre s'ouvre où il vous est demandé de taper le nom du fichier de données. En ce qui me concerne, il s'agit de "BoophilusAdultsDataIsoldistFarm.txt". Si tout se passe bien, il vous demande de cliquer sur la touche "Return" ou "Entrée" en français. Il vous faut ensuite choisir le

```
Data·R·microplus·pour·isoldist¶
B12¶
C07¶
D12¶
D10¶
A12¶
C03¶
Pop¶
16603.91667»    -2186.444444»    ,»    192194»    145192»    09:
16603.91667»    -2186.444444»    ,»    200200»    179179»    10(
16603.91667»    -2186.444444»    ,»    192194»    179192»    09:
16603.91667»    -2186.444444»    ,»    192198»    145192»    09:
16603.91667»    -2186.444444»    ,»    192194»    179179»    09:
16603.91667»    -2186.444444»    ,»    198200»    145179»    11:
16603.91667»    -2186.444444»    ,»    194198»    179179»    10(
16603.91667»    -2186.444444»    ,»    194198»    145192»    09:
16603.91667»    -2186.444444»    ,»    194198»    145145»    09:
16603.91667»    -2186.444444»    ,»    194198»    179179»    11:
16603.91667»    -2186.444444»    ,»    194198»    145179»    10(
16603.91667»    -2186.444444»    ,»    198200»    145192»    09:
16603.91667»    -2186.444444»    ,»    194198»    179192»    09:
16603.91667»    -2186.444444»    ,»    194198»    145178»    09:
16603.91667»    -2186.444444»    ,»    192198»    145192»    09:
16603.91667»    -2186.444444»    ,»    194198»    179179»    10(
16603.91667»    -2186.444444»    ,»    198198»    145145»    09:
16603.91667»    -2186.444444»    ,»    194198»    150192»    10·
16603.91667»    -2186.444444»    ,»    194198»    145150»    11:
16603.91667»    -2186.444444»    ,»    194194»    145145»    09:
16603.91667»    -2186.444444»    ,»    194198»    179192»    09:
16603.91667»    -2186.444444»    ,»    194194»    145192»    09:
16603.91667»    -2186.444444»    ,»    194198»    145192»    09:
16603.91667»    -2186.444444»    ,»    194198»    000000»    09:
16603.91667»    -2186.444444»    ,»    194198»    145179»    09:
16603.91667»    -2186.444444»    ,»    194198»    150150»    11:
16603.91667»    -2186.444444»    ,»    194200»    179179»    10:
Pop¶
16548.02778»    -2156.722222»    ,»    198198»    179179»    11:
16548.02778»    -2156.722222»    ,»    194198»    179192»    10:
16548.02778»    -2156.722222»    ,»    194200»    150179»    09:
16548.02778»    -2156.722222»    ,»    194194»    145179»    09:
16548.02778»    -2156.722222»    ,»    192196»    000000»    09:
16548.02778»    -2156.722222»    ,»    194198»    145145»    09:
16548.02778»    -2156.722222»    ,»    194194»    000000»    09:
16548.02778»    -2156.722222»    ,»    194198»    150185»    09:
16548.02778»    -2156.722222»    ,»    198198»    150179»    09:
```

Figure 76
Données pour Genepop 4 avec la ligne de titre (sans virgule),
les loci puis les génotypes précédés de leurs coordonnées GPS, suivies d'une virgule.
Chaque ferme est séparée par un "Pop" et il ne doit pas rester de colonne ou de ligne vide.

menu 6 puis le sous-menu 6. Nous sommes en deux dimensions, donc il faut choisir le logarithme naturel des distances géographiques. Tapez donc "l". On vous demande la distance minimale. Comme cela n'a pas beaucoup d'importance, ainsi que nous l'avons déjà vu, et que de toutes façon le test de Mantel n'en tiendra pas compte, tapez une toute petite valeur (0 étant exclu à cause de la transformation log). J'ai pour ma part tapé 0,01 puis "Entrée". Pour le nombre d'itérations de la chaîne de Markhov tapez 1 000 000.

Genepop génère plusieurs fichiers. Le premier à regarder (sinon le seul) est celui portant l'extension "iso". On y voit que la régression de pente b = 0,00362341 est marginalement significative selon le test de Mantel (P-value = 0,066394). Cette pente est cependant significativement différente de 0 selon l'intervalle de confiance issu de bootstraps IC 95 % = [0,00039310987 ; 0,0078657635]. Ce résultat est en fait similaire à ce qui avait été trouvé dans Koffi *et al.* (2006a) (b = 0,00054) ou

De Meeûs *et al.* (2010) (b = 0,0017) qui n'avaient pas utilisé tout à fait les mêmes données qu'ici et avaient considéré chaque infra-population séparément pour gagner en puissance (les P-values deviennent en effet très significatives). Nous allons garder nos valeurs puisque nous savons qu'isolement par la distance il y a et que cela ne changera de toutes manières pas grand-chose. Ici, nous avons l'illustration de la décision statistique que doit toujours prendre le biologiste. Ici, le Mantel n'est pas significatif au seuil 5 %. Mais une étude plus approfondie contredit cela et dans ce cas, la moins mauvaise solution est de décider qu'il y a en effet isolement par la distance. Vous pourrez vérifier par vous-même qu'en séparant les infra-populations

```
BoophilusAdultsDataCattleFarm¶
B12¶
C07¶
D12¶
D10¶
A12¶
C03¶
Pop¶
Boul·,·192194·145192·092112·153153·195197·146152¶
Boul·,·200200·179179·108112·151152·196196·152154¶
Boul·,·192194·179192·092112·153153·197199·152154¶
Boul·,·192198·145192·092112·152153·195196·146158¶
Boul·,·192194·179179·092100·152152·195199·152154¶
Boul·,·198200·145179·112118·152153·196197·146154¶
Boul·,·194198·179179·108112·152153·196197·146154¶
Boul·,·194198·145192·092100·153154·196197·146168¶
Boul·,·194198·145145·092102·152153·197199·154154¶
Boul·,·194198·179179·112112·152153·196197·154168¶
Boul·,·194198·145179·108112·153153·196197·154154¶
Boul·,·198200·145192·092092·153154·196197·154154¶
Boul·,·194198·179192·092112·154155·197199·146154¶
Boul·,·194198·145178·092112·153154·196197·168168¶
Boul·,·192198·145192·092100·154154·197199·146154¶
Boul·,·194198·179179·108112·153153·195196·152154¶
Boul·,·198198·145145·092102·152153·197197·152158¶
Boul·,·194198·150192·104112·152154·195197·146154¶
Boul·,·194198·145150·112112·152152·196199·146154¶
Boul·,·194194·145145·092112·152153·195197·154168¶
Boul·,·194198·179192·092092·152154·197197·146168¶
Boul·,·194194·145192·092112·152152·197199·154154¶
Boul·,·194198·145192·092112·152154·196197·152154¶
Boul·,·194198·000000·092112·152155·000000·154154¶
Boul·,·194198·145179·092112·152154·197199·146146¶
Boul·,·194198·150150·112112·152152·196199·146154¶
Boul·,·194200·179179·102102·152153·197199·154158¶
Pop¶
Bour·,·198198·179179·112112·152153·195197·146168¶
Bour·,·194198·179192·102102·153155·197199·146168¶
Bour·,·194200·150179·092102·152152·197199·146154¶
Bour·,·194194·145179·092100·152154·196199·154154¶
Bour·,·192196·000000·092100·151154·196199·154154¶
Bour·,·194198·145145·092092·151154·197199·152154¶
```

Figure 77
Extrait du jeu de données des génotypes microsatellites des tiques *Rhipicephalus microplus* au format Genepop pour Genetix, LDNe et Estim.

de tiques, le test devient très significatif. Puisque la pente est connue, nous pouvons donc calculer le voisinage qui est de $N_b = 1/b = 4\pi D\sigma^2 = 275{,}98$ individus. Et donc le produit de la densité efficace par la surface de dispersion $D\sigma^2 = 21{,}96$. Il nous faudrait maintenant une estimation indépendante de la densité des tiques sur parcelles pour pouvoir estimer la distance moyenne séparant des adultes reproducteurs de leurs parents.

EFFECTIFS EFFICACES

Ici trois méthodes sont disponibles : la méthode de Balloux (2004) sur les F_{IS}, la méthode de Waples et Do (2008) (en principe plus fiable que la méthode de Bartley et plus commode à implémenter) basée sur les déséquilibres de liaison et celle de Vitalis et Couvet (2001a-c) basée sur les corrélations alléliques intra et inter loci. Pour les trois méthodes, nous allons utiliser le fichier complet avec un sous-échantillon par ferme sous un format Genepop (extension .gen) comme dans la figure 77.

Pour estimer les F_{IS} par sous-échantillon avec leur bootstrap, nous allons utiliser Genetix (Belkhir *et al.*, 2004) qui offre une procédure directe par menu déroulant. Ouvrez Genetix, allez dans le menu "Fichier" puis "Importer". Cliquez dans le bouton "Genepop" et tapez "*.gen" dans la case "Nom du fichier", comme indiqué en figure 78.

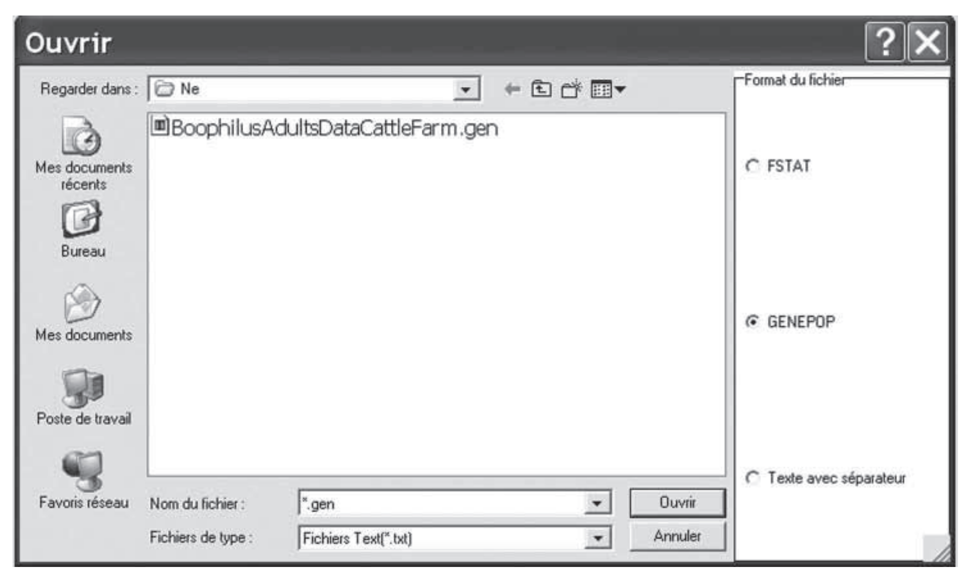


Figure 78
Menu Genetix pour importer le fichier des données microsatellites de *Rhipicephalus microplus* au format Genepop.

Figure 79
Sélection de l'option de calcul de bootstrap sur les F_{IS} par sous-échantillon sous Genetix.

Le fichier apparaît alors dans le cadre. Cliquez deux fois dessus et il s'ouvre sous Genetix. Sélectionnez le menu "Fstats", "Test sur Fis" et Bootstrap sur Fis par pop."[15], comme indiqué dans la figure 79.

Un menu s'ouvre où vous n'avez que deux choses à faire. Augmentez le nombre de bootstraps (en ce qui me concerne 10 000 j'aime bien), et cliquez ensuite sur "OK". Le résultat est disponible dans un fichier *.fis. Il faut ensuite appliquer la formule de l'équation 12 de Balloux (2004) :

$$N_e = \frac{-1}{2F_{IS}} - \frac{F_{IS}}{1+F_{IS}}$$

et de taper "Infinity" pour les valeurs négatives (quand le $F_{IS} > 0$).

Étant donné la tendance aux déficits en hétérozygotes, peu de valeurs exploitables ressortent de cette analyse (un seul $N_e = 6$ pour Bouloupari), mais on peut estimer que la limite inférieure (à 95 %) des N_e est en moyenne de 208 individus.

Pour la méthode de Waples et Do (2008), lancez LDNe. Cliquez sur le bouton "Search" et allez chercher votre fichier. Sélectionnez votre fichier et cliquez sur le bouton "OK" puis sur "Run LDNe". Attention, prenez garde que le fichier ne soit pas resté ouvert dans une autre application, auquel cas LDNe ne produit qu'un fichier de résultat *PL3.out vide. Sinon, après un travail rapide dans une fenêtre DOS, les calculs sont disponibles dans ce fichier. Nous allons prendre les valeurs calculées avec tous les allèles de fréquences au moins égales à 0,01 (valeurs les plus à droite) et prendre

[15] Je me suis aperçu sur le tard que les bootstraps de Genetix se font ici sur individus et non sur loci, ce qui peut poser des problèmes, surtout dans les petits échantillons (risque de rééchantillonner trop de fois le même individu) (je ne sais pas pourquoi les auteurs ont préféré cette option hétérodoxe). Ici, ça ne change rien eu égard aux résultats obtenus.

l'intervalle de confiance de jackknife. Nous obtenons beaucoup plus de valeurs utilisables avec un N_e moyen de 380 avec un intervalle de confiance moyen de [93, 440].

Pour la méthode de Vitalis et Couvet, nous allons ouvrir Estim, cliquer sur "File", "Open" et sélectionner le fichier. Ensuite, nous allons cliquer sur "Analysis", "Identity measures". Cliquez sur "Save" et choisissez un nom du genre "BooNeEstimRes.txt" pour l'enregistrer. Retournez au menu "Analysis" et cliquez maintenant sur "Ne inferences" et une fois encore sur "Save" en gardant le même nom. Ignorez l'avertissement en cliquant sur "Oui". Ici, très peu de résultats utilisables sont disponibles (comme d'habitude avec Estim) et seul Bouloupari donne un N_e = 1 429 et un taux de migration de 0,007 (il s'agit d'une estimation pour un modèle en île, donc une sorte de moyenne de ce qui vient de partout).

DENSITÉ EFFICACE ET DISTANCE DE DISPERSION PARENTS-DESCENDANTS ADULTES

En prenant la moyenne des différentes valeurs obtenues sur l'ensemble des méthodes, on obtient N_e = 605. La surface d'une exploitation est en moyenne d'environ 3 km² (Barré, communication personnelle). La densité efficace devient donc D_e = 202 tiques par km². En utilisant la taille de voisinage calculée plus haut, ainsi que son intervalle de confiance de Bootstrap à 95 %, nous obtenons une dispersion entre adultes reproducteurs et leurs parents de σ = 0,33 km [0,22 ; 1]. En utilisant le modèle de Rousset (1997), on obtient une estimation du taux de migration entre dèmes adjascents de $m = 2D\sigma^2/N_e$ = 0,07.

RECHERCHE DE LA SIGNATURE D'UN GOULOT D'ÉTRANGLEMENT

Le logiciel Bottleneck (Piry *et al.*, 1999) (voir aussi Cornuet et Luikart, 1996), que vous pouvez télécharger gratuitement à http://www.montpellier.inra.fr/URLB/bottleneck/bottleneck.html, utilise des fichiers au format Genepop et implémente son algorithme dans chaque sous-échantillon (fermes) identifié. Nous allons donc réutiliser le fichier "BoophilusAdultsDataCattleFarm.gen" (le même que pour les analyses LDNe et Estim).

Lancez Bottleneck. Laissez la photo du martin pêcheur disparaître (quelques secondes). Un panneau apparaît tel qu'en figure 80. Cliquez sur le bouton "Add data file..." et allez chercher votre fichier dans le menu qui apparaît. N'hésitez pas à taper *.gen dans la case "File name" ou "Nom du fichier" pour trouver les fichiers avec extension .gen. Ensuite, cochez le carré "T.P.M." et décochez les carrés "sign test "et "standardized differences test". En effet, il est intéressant de regarder ce qui se passe aussi en faisant l'hypothèse d'un modèle de mutation en deux phases (*two phases model* en anglais, TPM). Il est montré que, si un goulot d'étranglement a réellement eu lieu, on le détectera très fortement avec l'hypothèse IAM, moyennement avec le TPM et faiblement avec le SMM (Cornuet et Luikart, 1996), alors qu'en cas d'absence de goulot d'étranglement mais en population structurée en petites sous-populations, on pourra détecter faussement une signature de goulot d'étranglement avec IAM, mais exceptionnellement (voir jamais) avec TPM et jamais avec SMM (De Garine-Wichatitsky *et al.*, 2009 ; De Meeûs *et al.*, 2010). Donc, en cas de tests très significatifs pour les trois procédures, on peut être assez confiant. Le test le plus puissant et robuste pour tester un goulot d'étranglement est le Wilcoxon (Cornuet et Luikart, 1996) donc autant ne pas s'embarrasser avec les deux autres. Faites attention à ce que votre fichier soit au bon format (en particulier, pas de colonne ni de ligne supplémentaire à la fin), sinon Bottleneck risque de se fermer sans prévenir. Pour les paramètres du TPM, je laisse les paramètres par défaut, car s'il fallait en choisir on n'en sortirait pas (infinité de combinaisons). On a donc 70 % des mutations de type SMM et 30 % qui impliquent l'ajout ou le retrait de plus d'un motif microsatellite avec une variance de 30. Vous pouvez cliquer sur "GO !" et laisser le logiciel travailler bien gentiment.

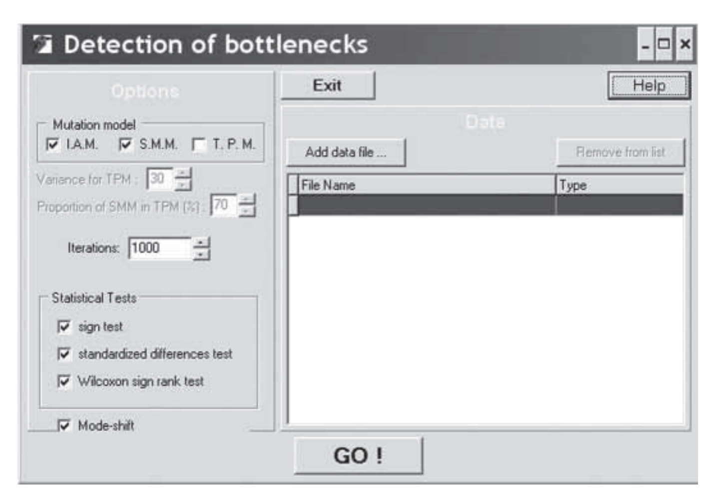


Figure 80
Panneau d'entrée de Bottleneck.

Quand c'est fini (au bout d'un petit quart d'heure sur ma machine), cliquez sur le bouton "Save results as text file" et nommez le fichier de résultats et enregistrez-le sous son nom, cliquez ensuite sur "Close" puis sur "Exit". Ce qui vous intéresse dans le fichier résultat correspond aux lignes "one tail for H excess" pour IAM, TPM et SMM. En effet, en cas de goulot d'étranglement récent, il est montré que la perte d'allèles se fait plus vite que la baisse de diversité génétique (H_s de Nei). Il en résulte que la diversité génétique observée sera plus grande que celle attendue eu égard au faible nombre d'allèles maintenus, si ce nombre d'allèles reflétait un équilibre entre mutation et dérive. Le logiciel explore, compte tenu de la distribution des allèles à chaque locus, cette diversité attendue si on était à l'équilibre (les auteurs l'appellent H_{eq}) et compare la valeur ainsi estimée de ce paramètre avec la diversité génétique (qu'ils notent H_e). *A priori*, il n'est pas nécessaire de s'intéresser aux autres tests. Les résultats figurent dans le tableau 33.

Dans ce tableau nous constatons que le signal est fort puisque significatif partout pour IAM et TPM, mais cependant nulle part pour SMM. Pour obtenir des *P*-values globales sur l'ensemble des fermes, nous allons utiliser le test binomial généralisé de Teriokhin *et al.* (2007) implémenté dans MultiTest V 1.2 (De Meeûs *et al.*, 2009). Il y a huit tests et donc k = 8. Pour IAM cela va vite, car toutes les *P*-values = 0,00781. On pose directement 0,0001 pour α, on clique sur "Test for k' and look for alpha'", on laisse k' à k/2 = 4 (recommandé) et on clique sur "Go!". Le test renvoie un seuil de 0,0355 qui est très supérieur à 0,00785. Pour IAM la *P*-value combinée est donc inférieure à 0,0001. J'estime en effet que des valeurs inférieures n'ont pas de sens en génétique des populations naturelles et c'est pourquoi je ne descends

Tableau 33
Résultat des tests de signature de goulot d'étranglement récent chez les tiques *Rhipicephalus microplus* dans les différents élevages échantillonnés en Nouvelle-Calédonie. Les *P*-values correspondent aux tests de Wilcoxon unilatéraux.

Ferme	IAM	TPM	SMM
Bouloupari	0,00781	0,01563	0,21875
Bourail	0,00781	0,00781	0,57813
Canala	0,00781	0,03906	0,71875
Gadji	0,00781	0,01563	0,71875
La Foa	0,00781	0,02344	0,65625
Poquereux	0,00781	0,01563	0,57813
Port-Laguerre	0,00781	0,01563	0,42188
Sarramea	0,00781	0,02344	0,21875

jamais en dessous de 0,0001. Pour le TPM, la quatrième plus petite *P*-value = 0,01563 est toujours inférieure à 0,0355. Ici aussi, la *P*-value combinée est inférieure à 0,0001. Pour le SMM, la quatrième plus petite *P*-value est de 0,57813. Or la valeur maximale pour α' est 0,5. Il est clair que pour SMM, la *P*-value est > 0,5 (on peut noter 0,57813 pour donner un chiffre comme dans De Meeûs *et al.*, 2009).

La conclusion, eu égard aux niveaux de significativité obtenus avec le IAM et le TPM, est qu'il existe bien une signature de goulot d'étranglement dans les fermes. Ce goulot correspond le plus vraisemblablement à l'introduction accidentelle de quelques individus *R. microplus* en 1942. Conformément au modèle de Cornuet et Luikart (1996) (voir plus haut en p. 220), compte tenu du nombre de loci, la détection de ce goulot d'étranglement suppose alors que l'effectif efficace post-goulot d'étranglement (de la première ferme touchée) a été de N_{eb} = [49, 1 220], soit une gamme de valeur remarquablement convergente avec la gamme donnée par les autres méthodes d'estimation de N_e.

CONCLUSIONS

Nos analyses ont permis de montrer que l'unité démographique de *R. microplus* n'est pas l'individu hôte (avec son infra-population) comme pressenti, mais plutôt l'élevage ou troupeau d'une ferme. Cette tique passe donc, du stade larve à adultes, librement d'une bête à l'autre d'un troupeau et est donc parfaitement susceptible de propager des maladies telles que l'anaplasmose si cette dernière était introduite sur l'île.

Il apparaît que les populations locales de *R. microplus* (troupeau) sont structurées en fratries, ce qui suppose une réussite hétérogène entre pontes, compatible avec les traitements acaricides réguliers : la ponte des femelles tombées au sol juste avant traitement n'est pas affectée, les autres disparaissent presque toutes. Cette structure génétique particulière est accompagnée d'une légère signature d'appariement assorti qui peut très bien en être une conséquence : les membres d'une même fratrie étant plus synchrones ensemble qu'avec les autres. Ceci explique les légers déficits en hétérozygotes significatifs observés.

Il existe un isolement par la distance dont le modèle nous permet d'inférer un voisinage de taille 276 individus, notion particulièrement difficile à comprendre s'il en est, mais qui permet d'estimer la surface de dispersion entre adultes et les parents leur ayant donné naissance. Cette dernière s'avère relativement modeste avec un rayon de l'ordre des 300 m par génération (entre 200 m et 1 km), soit au plus 1,2 km par an en moyenne (si quatre générations par an et pas de retour en arrière). Cette dispersion découle de l'estimation de densités efficaces relativement importantes

d'environ 200 tiques/km², soit 600 tiques « reproductrices » par élevage. Compte tenu du fait que nos estimations d'effectifs efficaces sont probablement sous-évaluées (voir Bouyer *et al.*, 2009), que les déficits en hétérozygotes témoignent d'effectifs efficaces inférieurs aux effectifs réels, on se retrouve avec des densités de tiques importantes (plus de 1 000/km²) telles qu'observées sur le terrain (Koffi *et al.*, 2006a), malgré les traitements acaricides. Ces derniers semblent donc d'un impact léger sur la démographie de la tique. Si nous considérons que la rotation des bêtes se fait sur 2 à 5 parcelles par génération de tiques (Koffi *et al.*, 2006a) et que chaque parcelle fait en moyenne 3 km², on peut en déduire que les tiques circulent sur une surface totale 6 à 15 km², soit (en considérant qu'il s'agit d'un disque de surface πr^2) sur un rayon de 0,8 à 1,1 km, donc dans le même ordre de grandeur que ce que la génétique semble indiquer. Il y a donc convergence remarquable entre observations directes et inférences par outil de génétique des populations. De grandes populations et un isolement important entre elles doivent favoriser l'apparition et l'installation rapide de mutations favorables et conférer un potentiel évolutif important à *R. microplus* (voir à ce titre Chevillon *et al.*, 2007b ; De Meeûs *et al.*, 2010). L'introduction unique à partir de peu d'individus en 1942 est compatible avec la signature d'un goulot d'étranglement assez fort. En fait, selon la figure 3A de Cornuet et Luikart (1996), avec moins de 10 loci, une moyenne de 170 allèles génotypés (85 individus) par sous-échantillon et 100 % de détection en IAM, cette détection n'est possible que si le goulot d'étranglement s'est fait avec un rapport taille de population avant/taille de population après α = [100 ; 1 000] et un paramètre τ = [0,25 ; 1]. Avec 244 générations, nous obtenons un effectif post-bottleneck $N_{eb} = t/2\tau$ = [122 ; 488], ce qui converge bien avec les autres résultats. Si on considère que la population d'origine des premières *R. microplus* colonisatrices avait une taille sensiblement équivalente à celle des N_{eb} trouvés en Nouvelle-Calédonie, on peut inférer que ce nombre $N_{intro} = N_{eb}/\alpha$ = [1 ; 5] tiques, c'est-à-dire à partir d'excessivement peu d'individus reproducteurs. Il est probable qu'il s'agit d'une introduction unique, car sinon plus difficile à détecter génétiquement, et donc que les dispositifs de restriction mis en place sur l'île ont été efficaces jusqu'à présent.

Il est clair que la qualité des loci utilisés (pas d'allèle drop out), au nombre de six seulement, ainsi que celle de l'échantillonnage ont seules permis d'aller aussi loin dans nos investigations, voir même beaucoup plus loin si on se réfère aux autres travaux associés à ce projet non abordés dans ce manuel (De Meeûs *et al.*, 2010).

7 Génétique des populations de *Trypanosoma brucei gambiense* en Afrique de l'Ouest

INTRODUCTION

Le jeu de données que nous allons analyser a fait l'objet d'une publication en 2009 (Koffi *et al.*, 2009). Il va nous permettre d'explorer comment adapter les outils de la génétique des populations aux organismes à reproduction majoritairement asexuée.

ÉTAT DES LIEUX

Les trypanosomiases africaines sont des maladies à vecteur transmises normalement par des glossines (mouches tsé-tsé) et parfois mécaniquement par d'autres insectes piqueurs (tabanides) ou même sexuellement pour *Trypanosoma equiperdum* (Brun *et al.*, 1998). La maladie du sommeil ou trypanosomiase humaine africaine (THA) est connue sous deux formes : la forme chronique, rencontrée en Afrique de l'Ouest et centrale, et la forme aiguë, qui sévit en Afrique de l'Est. La forme chronique de la THA est provoquée par *Trypanosoma brucei gambiense* type 1 (Tbg1) et représente plus de 90 % des cas recensés par l'Organisation mondiale de la santé (OMS) (WHO, 2006b). Une personne infectée par Tbg1 peut rester asymptomatique durant des années avant de déclarer la forme neurologique (dramatiquement spectaculaire) de la maladie. La forme aiguë de la THA est provoquée par *Trypanosoma brucei rhodesiense* (Tbr) pour laquelle les premiers symptômes neurologiques peuvent apparaître au bout de quelques semaines seulement. Ce schéma idéal n'est pas toujours très clairement suivi *in situ* et de nombreux variants cliniques sont trouvés pour les deux formes en conséquence de facteurs liés à l'hôte, au parasite, à l'environnement socio-économique ou écologique, voire même une combinaison de tous ces paramètres ou d'une partie d'entre eux (MacLean *et al.*, 2007). Sans traitement, les deux formes de la THA conduisent à une issue fatale (Garcia *et al.*, 2006 ; WHO, 2006b), bien que des enquêtes épidémiologiques suggèrent de plus en plus l'existence de porteurs sains capables de contrôler l'infection, voire même de la juguler (Garcia *et al.*, 2006). Après la flambée du début du XXe siècle, la THA semblait largement éradiquée dans le courant des années 1960. Elle a cependant réémergé dans les années 1980 en corollaire d'une baisse significative de la surveillance, de déplacements de populations, de guerres et de catastrophes naturelles (Aksoy *et al.*

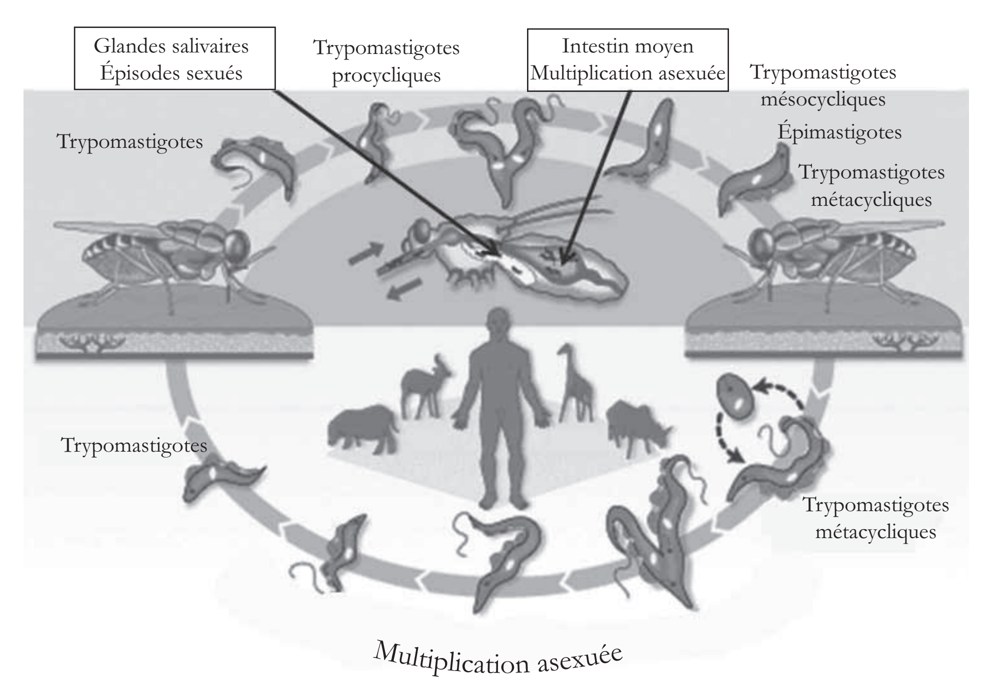


Figure 81
Le cycle de *Trypanosoma brucei*. La tsé-tsé injecte à l'hôte des trypomastigotes métacycliques lors d'un repas sanguin qui se transforment en stades trypomastigotes sanguins. Après une phase de multiplication asexuée, les trypomastigotes raccourcissent et peuvent alors être ingérés par une nouvelle tsé-tsé lors d'un nouveau repas sanguin sur l'hôte. Dans l'intestin moyen de la glossine, les trypomastigotes se transforment en trypomastigotes procycliques qui se multiplient par fission binaire. Dans l'intestin moyen antérieur, les trypomastigotes procycliques se transforment en trypomastigotes mésocycliques qui migrent alors dans les glandes salivaires où ils se transforment en épimastigote puis enfin en trypomastigotes métacycliques de nouveau.
Schéma inspiré d'une figure du TDR Wellcome/Trust (http://www.who.int/tdr/diseases/tryp/lifecycle.htm).

2005 ; Garcia *et al.*, 2006). En 2000, il a été estimé qu'environ 300 000 personnes étaient infectées et que seulement 10 à 15 % des 60 millions de personnes vivant dans les zones à risque étaient sous surveillance médicale (Garcia *et al.*, 2006). Grâce aux mesures de contrôle, il semble que nous soyons aujourd'hui dans un contexte d'élimination (Jamonneau, communication personnelle).

La trypanosomiase animale africaine (TAA ou nagana) est causée par différentes espèces de trypanosomes, classiquement : *T. brucei brucei* (Tbb), *T. congolense* (Tc) et *T. vivax* (Tv) qui affectent gravement la santé du bétail. La TAA représente un frein majeur au développement en Afrique subsaharienne et son coût annuel a été estimé à hauteur de 4,75 milliards de dollars américains (FAO, 2000 ; Bouyer *et al.*, 2009).

Trypanosoma brucei sl requiert deux hôtes séquentiels pour accomplir son cycle (fig. 81). Un vertébré (l'homme, par exemple), où le parasite se propage par fission binaire (clonalité), et une glossine (le vecteur) où une phase de propagation clonale dans l'intestin moyen précède une éventuelle sexualité, de type classique (meïose avec ségrégation, recombinaison et amphimixie) qui a lieu dans les glandes salivaires de la mouche tsé-tsé (MacLeod *et al.*, 2005a, b, c, 2006 ; Tait *et al.*, 2007).

En théorie, la recombinaison sexuée peut intervenir chez n'importe laquelle des espèces (ou sous-espèce, on y reviendra) du complexe *T. brucei* (Tait *et al.*, 2007). Il semblerait cependant que ceci ne concerne que les *T. brucei* d'animaux (i.e. Tbb), alors que la sexualité serait rare ou absente chez les souches infectant l'homme (Tbg1 et Tbr) (MacLeod *et al.*, 2000 ; De Meeûs et Balloux, 2005 ; Koffi *et al.*, 2009 ; Simo *et al.*, 2010). Cependant, ces inférences sont toujours l'objet de contestations, car elles dépendent fortement de la stratégie d'échantillonnage et notamment de ce qui est considéré comme appartenant ou non à la même espèce (Maynard-Smith *et al.*, 1993 ; MacLeod *et al.*, 2000). Par ailleurs, la plupart des investigateurs considèrent les déséquilibres de liaison comme des outils privilégiés de mesure de la clonalité, alors qu'il a été montré que ces déséquilibres de liaison sont très difficiles à estimer et dépendent fortement de la structure des populations cibles (De Meeûs et Balloux, 2004 ; Prugnolle et De Meeûs, 2010). Or les organismes tels que les trypanosomes ont de fortes chances de montrer des structures de populations assez cloisonnées. Pour les espèces diploïdes, comme c'est le cas des trypanosomes, le paramètre F_{IS} de Wright (Wright, 1965), qui mesure comme on l'a vu l'homozygotie des individus relative à l'homogénéité génétique de la sous-population dont ils sont issus, représente un outil beaucoup plus performant (De Meeûs et Balloux, 2005 ; De Meeûs *et al.*, 2006).

Un autre problème, spécifique à Tbg1, concerne la méthode d'isolement des souches. Il a en effet été montré que les profils enzymatiques de souches provenant du même patient, mais isolées par différentes méthodes, étaient différents. De là, l'idée que ces méthodes sélectionnaient des souches de parasites particulières (Jamonneau *et al.*, 2003), ce qui est gênant si on ne peut pas être certain d'avoir des échantillons représentatifs de la diversité présente. Ces méthodes d'isolement sont au nombre de trois : l'inoculation de rongeurs de laboratoire (IR) par du sang contaminé (trypomastigotes sanguins), peu efficace eu égard au manque de virulence des Tbg1 chez les rongeurs (Jamonneau *et al.*, 2003) ; la culture *in vitro* avec le kit d'isolation *in vitro* (KIVI) beaucoup plus efficace (Jamonneau *et al.*, 2003) ou enfin à partir des liquides biologiques (sang, lymphe des ganglions ou liquide céphalorachidien) directement. Ici, ce sont des extraits directs de sang ou BS (blood samples) qui ont été comparés aux deux autres.

Dans ce chapitre, nous allons revisiter pas à pas les données de l'article de Koffi *et al.* (2009) afin d'explorer le système de reproduction de ce pathogène, tester le biais occasionné par les différentes méthodes d'isolement, estimer la taille de ses



Figure 82
Localisation géographique des foyers de THA étudiés (marqués d'une étoile).

populations dans chaque foyer et le nombre de migrants sur un échantillon de 90 souches prélevées en Côte d'Ivoire dans le foyer de Bonon et en Guinée dans les foyers de Boffa et Dubréka (fig. 82), sur une période allant de 1998 à 2004.

LE JEU DE DONNÉES BRUTES

Les informations générales concernant les données sont présentées dans le tableau 34.

Les données brutes sont contenues dans le fichier "TrypanoBruceiTotDataGPS.txt" qui, en plus des données des 90 isolats cités plus haut, donne les génotypes d'un certain nombre de souches de référence de Tbg1, de Tbb, de Tbg2 (des Tbb trouvés chez l'homme en Côte d'Ivoire (Gibson, 2007) et de Tbr. Les données se présentent comme suit (fig. 83).

Tableau 34
Nombre d'isolats ($N_{isolats}$) de *Trypanosoma brucei gambiense* échantillonnés dans les différents foyers et années de l'étude. La surface occupée, la taille de la population humaine, les prévalences et le nombre présumé de personnes infectées (Prévalence×Population) sont également indiqués.

Pays	Foyer	Année	$N_{isolats}$	Surface (km²)	Population	Prévalence	$N_{infectés}$
Côte d'Ivoire	Bonon	2000	17	400	30 000	0,004	120
		2002	14				
		2004	17				
Guinée	Boffa	2002	20	2 400	25 000	0,0118	295
	Dubréka	1998	15	1 600	25 000	0,0075	187
		2002	7				

Nous avons besoin de rajouter une information manquante à ces données, les génotypes multilocus (MLGs), qui est une information extrêmement utile en génétique des populations clonales (Tibayrenc *et al.*, 1990 ; Tibayrenc *et al.*, 1991 ; Tibayrenc, 1998 ; 1999 ; Tibayrenc et Ayala, 2002 ; De Meeûs *et al.*, 2006). En ce qui me concerne, je le fais sous Excel. Je charge le fichier sous Excel. Je crée une

Long»	Lat»	Isolats»	Pays»	Foyer»	Année»	Isolement»	mic-bg1»	mic-
-4.943694444»	9.12»	B3/1/3·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162192»	1702
-4.944527778»	9.11625»	G10/6/2·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.950194444»	9.120555556»	S3/4/1·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.942111111»	9.128833333»	T66/4/2·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.934055556»	9.089694444»	402/1·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.955611111»	9.124777778»	B12/2/8·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.947472222»	9.117222222»	DF1/4·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.944361111»	9.120611111»	F41/7/2·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.949611111»	9.120583333»	F5/10M·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.944166667»	9.114444444»	G11/8/2·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.949666667»	9.128027778»	G17/6/1·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.989138889»	9.099666667»	S24/7/9·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.944055556»	9.114694444»	G3/10/25·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.954861111»	9.119416667»	F7/1/2·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.945166667»	9.114361111»	G11/6/4·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.944555556»	9.116944444»	S27/16/13·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-4.944694444»	9.117972222»	S27/2/6·KIVI·ms»	CI»	Bonon»	2000»	KIVI»	162194»	1702
-6.08978»	6.89971»	TT22/1·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
-6.08978»	6.89971»	TT22/1·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
6.04874»	6.92096»	S1/1/6·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
6.05518»	6.9203»	S7/2/2·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
-6.08978»	6.89971»	TT2/4·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.05518»	6.9203»	S7/2/2·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.04874»	6.92096»	S1/1/6·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.05326»	6.9219»	S12/9/5·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.04182»	6.91922»	T41/4/14·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.04182»	6.91922»	T41/4/14·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
6.05326»	6.9219»	S12/9/5·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
6.05438»	6.92206»	S14/5/1·KIVI·ms»	CI»	Bonon»	2002»	KIVI»	162194»	1702
6.05438»	6.92206»	S14/5/1·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1702
-6.08978»	6.89971»	TT2/4·RI·ms»	CI»	Bonon»	2002»	RI»	162194»	1742
»	»	B4/E120·BS·ms»	CI»	Bonon»	2004»	BS»	162194»	1702
»	»	B4/I245·BS·ms»	CI»	Bonon»	2004»	BS»	162194»	1702
»	»	B4/I36·BS·ms»	CI»	Bonon»	2004»	BS»	162194»	1702
»	»	B4/U163·BS·ms»	CI»	Bonon»	2004»	BS»	162194»	1702
»	»	B4/G13·BS·ms»	CI»	Bonon»	2004»	BS»	162194»	1702

Figure 83
Extrait du fichier de données de *Trypanosoma brucei*. En ligne figurent les différents isolats (comme d'habitude). Les deux premières colonnes donnent les coordonnées GPS des patients (pour Bonon 2000 seulement), suivent le nom de l'isolat, le pays, le foyer, la méthode d'isolement des souches et les huit loci microsatellites sur lesquels ces isolats ont été génotypés.

colonne “Somme” où je fais la somme de tous les allèles de tous les loci pour chaque isolat et je trie le tableau en fonction de “Somme”. Je crée une nouvelle colonne “Id” avec une fonction qui marque 1 quand, dans la colonne “Somme”, plusieurs chiffres qui se suivent sont égaux. Si la colonne “Somme” correspond à la colonne P de Excel, alors tapez “=SI(P3=P4;1;””)” en ligne 3 (deuxième isolat) de la colonne Id et copier cette cellule et la coller sur toutes celles du dessous. Je crée enfin une colonne “MLG” où je numérote dans l’ordre les génotypes en mettant le même chiffre pour ceux qui se répètent en m’aidant de la colonne “Id”, mais en prenant garde que l’identité de la somme résulte bien d’une identité multilocus. J’ai enregistré ce fichier sous le nom “TrypanoBruceiTotDataGPS.txt” où je vais ensuite supprimer les colonnes de calculs intermédiaires et ne garder que MLG en dernière colonne (après le dernier locus donc).

Il faut ensuite rendre ce fichier lisible par Create, ce qui nous permettra ensuite de le traduire pour n’importe quel logiciel. Par commodité, il convient de transformer d’abord tous les “0” en “000000”. Ensuite, il faut séparer les deux allèles de chaque locus en collant une colonne de tabulation entre les deux allèles de chaque locus (on obtient deux colonnes par locus donc). Il faut répéter donc sur la première ligne le nom des loci et faire en sorte que le nom de chaque locus ne dépasse pas six caractères (certains logiciels vont les tronquer sinon) et ne comporte pas de caractères spéciaux tels que – ou /. Enfin, certains loci ont des allèles dont la taille est inférieure à 100. Il faut penser à leur rajouter un 0 devant (par exemple, 085). Il faut coder aussi les MLGs avec trois caractères et dupliquer cette colonne (rendre ce “locus” diploide homozygote). Les MLGs seront utilisés pour des tests de randomisations d’individus entre sous-échantillons et pour mesurer l’indice de différenciation, soit θ l’estimateur du F_{ST}. Nous avons déjà vu que le F_{ST} ne dépend que de l’homogénéité interindividuelle dans et entre sous-populations, la diploïdisation homozygote n’a donc aucun effet à ce niveau.

En principe, nous pouvons commencer. Nous allons déjà nous débarrasser du facteur « technique d’isolement » afin, en cas de non-significativité, de pouvoir ignorer ce facteur et travailler sur de plus grands sous-échantillons.

TESTER L’EFFET DE LA TECHNIQUE D’ISOLEMENT DES SOUCHES

Création d’un fichier Fstat et MSA

Nous allons utiliser ici les procédures F_{ST} par paire de sous-échantillons et les tests de différenciation par paire de sous-échantillons sous Fstat et aussi créer un dendrogramme. Nous ne pouvons pas utiliser HierFstat ici car le facteur « technique d’isolement » est

un facteur croisé (ou orthogonal) et non pas hiérarchisé, comme cela est requis pour HierFstat (voir la discussion à ce sujet dans De Meeûs et Goudet, 2007). Il faut donc créer ce fichier avec par exemple Create (il s'agit juste d'une suggestion). N'oubliez pas de créer une nouvelle colonne qui informe sur le foyer, l'année et la méthode d'isolement (Bon00KI pour Bonon 2000 KIVI) et de trier selon cette colonne. Quand cela est en ordre, on lance Create pour convertir le fichier au format Fstat et MSA (qui nous servira à construire une matrice de distances génétiques). N'oubliez pas de supprimer la colonne supplémentaire inutile du fichier ".lab" que Create va créer. Vous pouvez également raccourcir les noms de fichiers à votre convenance.

Analyse Fstat par paire de sous-échantillons

Il faut charger ensuite le fichier .dat sous Fstat. Il faut sélectionner les loci (pas le locus MLG dans un premier temps) et les sous-échantillons pertinents (pas les souches de références ni les sous-échantillons où il n'y a eu qu'une seule méthode de prélèvement). Ceci se fait avec le menu déroulant "Options" de Fstat et les sous-menus "Label file for pops" pour indiquer le fichier contenant le nom des sous-échantillons (plus facile pour la suite), "Loci to use" (on sélectionne tout sauf MLG) et "Samples to use" (on sélectionne les sous-échantillons de Bonon en 2002 et 2004 qui sont les seuls où plusieurs méthodes de prélèvements sont disponibles). Dans le cadre principal du menu Fstat, cochez "Fst per pair of samples", "Pairwise tests of differentiation" et activez le bouton "1/1000" de "Nominal level to multiple tests" (pour avoir suffisamment de permutations). Enfin, cliquez sur "Run". Nommez le nouveau fichier (T-BruceiBetweenIsolationMetFstat.dat) (nous avons en effet sélectionné des loci et sous-échantillons particuliers pour ce test) et cliquez sur "Enregistrer" pour lancer l'analyse Fstat. Deux fichiers de sortie Fstat nous intéressent, celui qui possède les F_{ST} par paire de sous-échantillon et qui porte l'extension "fst" et celui qui donne les *P*-values avec l'extension "pvl". Les seules paires qui nous intéressent sont celles qui comparent deux méthodes dans un même sous-échantillon. Comme on le voit dans le tableau 35, nous obtenons quatre comparaisons qui toutes présentent un estimateur de $F_{ST} < 0$ non significatif.

Tableau 35
Résultats des mesures et tests de significativité par paire de méthodes d'isolement des souches de *Trypanosoma brucei gambiense* 1 à Bonon en 2002 et en 2004. Données avec les loci individuels.

Année	Méthode 1	Méthode 2	F_{ST}	*P*-value
2002	KIVI	Rodent inoculation	- 0,0164	0,9547
	Blood sample	KIVI	- 0,0088	0,6749
2004	Blood sample	Rodent inoculation	- 0,0181	0,8319
	KIVI	Rodent inoculation	- 0,0131	0,7192

Les organismes clonaux ont la fâcheuse habitude de générer une corrélation entre les loci (déséquilibres de liaison), d'où la présence de génotypes multilocus. Cela pourrait conduire un test de différenciation, par effet d'auto-corrélation, à pencher trop fort dans une direction ou l'autre (bien qu'ici les résultats soient peu ambigus). Pour valider notre test, l'utilisation des génotypes multilocus ou MLGs comme autant d'allèles d'un même et unique locus est une option efficace. Nous allons donc répéter ce que nous venons de faire, mais en ne gardant que le "locus" MLG. L'analyse du nouveau jeu de données ainsi créé (T-BruceiBetweenIsolationMetFstatMLG.dat) aboutit aux résultats présentés dans le tableau 36. On voit encore que la différenciation n'est pas significative avec des mesures de différenciation systématiquement négatives ou nulles.

Tableau 36
Résultats des mesures et tests de significativité par paire de méthodes d'isolement des souches de *T. brucei gambiense* 1 à Bonon en 2002 et en 2004. Données MLG.

Année	Méthode 1	Méthode 2	F_{ST}	*P*-value
2002	KIVI	Rodent inoculation	- 0,0399	0,9061
	Blood sample	KIVI	- 0,0256	1
2004	Blood sample	Rodent inoculation	- 0,0345	1
	KIVI	Rodent inoculation	0,0000	1

Analyse NJTree

Nous allons pour cela créer un fichier MSA avec Create. N'oubliez pas de retirer le locus MLG, ainsi que les souches de référence qui n'ont pas lieu d'être ici. Quand votre fichier est prêt, copiez-le dans le répertoire de MSA (ou copiez MSA dans votre répertoire de travail). Lancez MSA, tapez "i" pour choisir le nom de votre fichier de données et tapez le nom complet de ce fichier (celui que vous venez de créer avec Create). Tapez ensuite "d" pour le menu des distances, puis "p" pour choisir le type de distance. Ensuite, tapez "c" pour sélectionner le calcul par paire de sous-échantillons, puis les chiffres correspondant aux distances à sélectionner ou à désélectionner. En principe, on garde la distance harmonique de Cavalli-Sforza et Edwards (*chord distance*) qui est réputée produire les meilleurs NJTree, eux-mêmes réputés donner les arbres dotés de la meilleure topologie (Takezaki et Nei, 1996). Donc on va garder l'option correspondant à cette distance "on" (option 7, indissociable de l'option 8, pour une raison qui m'échappe). Tapez enfin "!" pour lancer les calculs. MSA crée un répertoire plein de sous-répertoires pleins de trucs inutiles. Intéressez-vous à ce qu'il y a dans le répertoire "Distance_data" dans le quel se trouve le fichier "CAS_Pop.txt" qui nous intéresse. Il faut ouvrir ce fichier avec un tableur ou un bon éditeur de texte. Il contient la matrice des distances harmoniques de Cavalli-Sforza et Edwards entre toutes les paires de sous-

```
#mega¶
!TITLE··Genetic·distance·data·from·T·brucei;¶
!Format·DataType=distance;¶
!Description¶
·····Between·Clusters·Cavalli-Sforza·and·Edwards·Chord·Distance¶
·····8·Microsatellite;¶
¶
#Bon00KI¶
#Bon02KI¶
#Bon02RI¶
#Bon04BS¶
#Bon04KI¶
#Bon04RI¶
#Bof02KI¶
#Dub02KI¶
#Dub98KI¶
¶
¶
0.17393¶
0.20502»    0.09508¶
0.26111»    0.21383»    0.23176¶
0.20953»    0.18884»    0.21322»    0.23967¶
0.21738»    0.16867»    0.22307»    0.22949»    0.15887¶
0.57362»    0.53260»    0.50151»    0.54230»    0.51877»    0.53714¶
0.45882»    0.43909»    0.38943»    0.42733»    0.41053»    0.43489»    0.22836¶
0.38105»    0.39050»    0.34547»    0.38294»    0.38359»    0.41611»    0.36047»    0.22948
```

Figure 84
Extrait du fichier de données de matrice de distances
pour fabriquer un NJTree sous Mega (les “>>” représentent des tabulations).

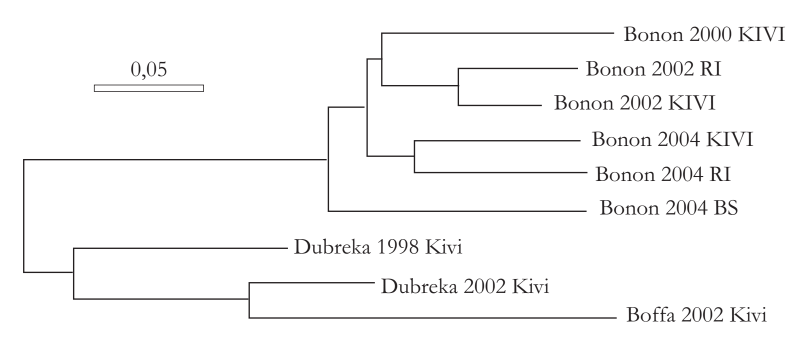


Figure 85
Résultat du NJTree basé sur la distance harmonique de Cavalli-Sforza
et Edwards entre paires de sous-échantillons calculée à partir de huit loci microsatellites.

échantillons. Il faut ensuite ouvrir un fichier type MEGA (Kumar *et al.*, 2004), comme décrit dans la figure 84. Le résultat obtenu est présenté en figure 85. On voit bien que la méthode d'isolement n'est pas un paramètre très important. Notez que le dendrogramme obtenu diffère de celui publié par Koffi *et al.* (2009), car ce dernier était basé sur six des loci (Micbg6 et Trbpa avaient été éliminés pour des raisons que nous verrons plus loin) et sur des distances évaluées par Genetix qui calcule en fait une autre distance que la distance harmonique de Cavalli-Sforzza et Edwards (contrairement à ce qui est dit). Mais la conclusion générale ne change pas.

Nous pouvons donc désormais ignorer le facteur méthode d'isolement dans les analyses qui vont suivre.

DÉSÉQUILIBRES DE LIAISON, HOMOZYGOTIE RELATIVE LOCALE ET SYSTÈME DE REPRODUCTION

Création du fichier Fstat

En reprenant le fichier de départ, nous allons construire un fichier Create où chaque combinaison de Foyer×Année d'isolement correspondra à une population différente et en éliminant pour le moment les souches de référence. Une fois que cela est fait, on traduit ce fichier au format Fstat en suivant la même procédure que précédemment.

Analyse des déséquilibres de liaison et des F_{IS}

J'ai appelé mon fichier "T-BruceiFoyAnCI&Guin.dat". Dans l'analyse Fstat, après avoir chargé ce fichier et choisi un fichier "Label for pops" dans "Options", j'ai coché les cases correspondant aux fréquences alléliques, mesures de diversité génétiques sur l'ensemble et par locus et population, le test sur le F_{IS} global et pour chaque locus dans chaque sous-population, ainsi que celui pour le déséquilibre de liaison dans chaque population et entre chaque paire de loci. Je m'arrange pour qu'il y ait 10 000 permutations au moins. Une fois que tout est prêt, il faut cliquer sur "Run" et attendre que toutes les permutations soient finies (1 mn chez moi). Les résultats apparaissent dans le fichier "T-BruceiFoyAnCI&Guin.out".

Déséquilibres de liaison

Nous ne regardons que les tests sur l'ensemble des sous-échantillons et par paire de loci. Sur les 21 tests possibles, 18 paires de loci sont significativement en déséquilibre de liaison au seuil 5 %. Cela est largement au-dessus des 5 % attendus (même pas besoin de faire un test binomial, mais bon cela donne une *P*-value < 0,0001). Treize tests restent significatifs au seuil Bonferroni séquentiel et chaque locus est impliqué au moins une fois dans une liaison significative à ce seuil. Nous pouvons conclure qu'une liaison statistique très significative existe entre tous les loci, c'est-à-dire que cette association concerne l'ensemble du génome des trypanosomes.

Excès d'hétérozygotes locaux

Globalement, il existe un important excès d'hétérozygotes avec un F_{IS} = - 0,611 et un intervalle de confiance à 95 % de [- 0,76, - 0,473]. Cet excès est très significatif (*P*-value < 0,0001). Globalement, Fstat ne teste que $F_{IS} > 0$, mais il suffit de prendre 1-*P*-value, qui est ici de 0,9999, ce qui donne 0,0001. Regardons ce qui se passe locus par locus. Dans le fichier "T-BruceiFoyAnCI&Guin.out", il s'agit maintenant

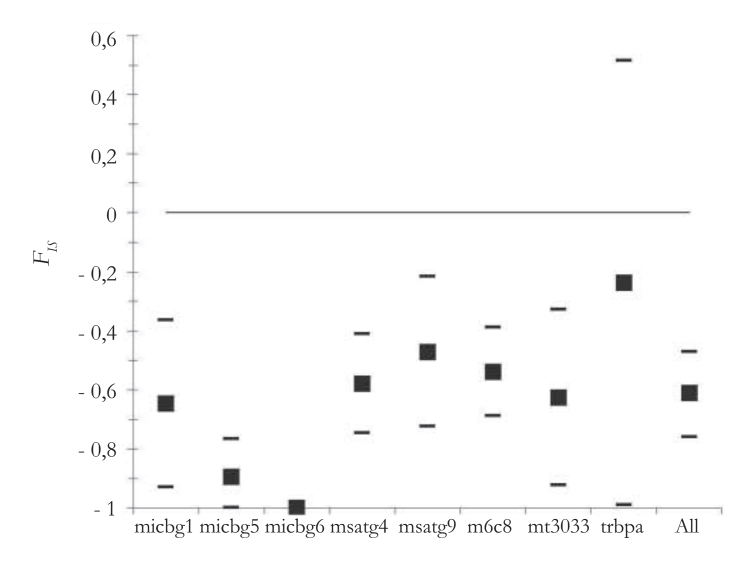


Figure 86
Valeurs de F_{IS} par locus et sur l'ensemble (All), intervalles de confiance à 95 % de jackknife sur les sous-échantillons (pour les loci) ou de bootstrap sur les loci (pour la moyenne globale : All).

de récupérer les valeurs de F_{IS} (smallf) par locus sur l'ensemble des sous-échantillons, leur erreur standard de jackknife (StrdErrFis) sur les sous-échantillons (*over populations*). Pour six sous-échantillons (donc 6 - 1 = 5 ddl), le paramètre $t \approx 2{,}57$ au seuil 5 % (cf. p. 72-74 de la 1re partie de ce manuel). Pour chaque locus, l'intervalle de confiance se calcule donc avec les formules F_{IS}-2,57xStrdErrFis pour la limite inférieure, qui ne peut dépasser - 1, et F_{IS}-+2,57xStrdErrFis pour la limite supérieure, qui ne doit pas dépasser + 1. Les valeurs d'intervalle de confiance qui dépassent les valeurs - 1 et + 1 doivent donc être artificiellement ramenées à ces valeurs frontières. En faisant cela, nous supposons que les F_{IS} suivent la loi normale, ce qui est sans doute faux. D'un autre côté, nous n'utiliserons pas ces intervalles de confiance pour une décision statistique, mais pour illustrer le comportement des différents loci dans un graphique. Nous pouvons ainsi réaliser le graphe de la figure 86. On notera que toutes les *P*-values = 0,0001 sauf pour trbpa (*P*-value = 0,0011). On peut aussi noter que deux loci sortent du lot, micbg6 qui est en fait fixé hétérozygote 182/266 dans tous les échantillons et trbpa dont la variance est anormalement élevée. Ce locus est d'ailleurs situé dans une zone codante (Roditi *et al.*, 1998) et nous avons là typiquement une bonne raison d'éliminer une source d'information qui apporte plus de confusion qu'autre chose.

Pour recommencer cette analyse sans le locus trbpa, il suffit de recharger le fichier dans Fstat et de sélectionner les sept autres loci. Fstat crée un autre fichier que j'ai personnellement nommé "T-BruceiFoyAnCI&Guin-CleanLoci.dat". En regardant

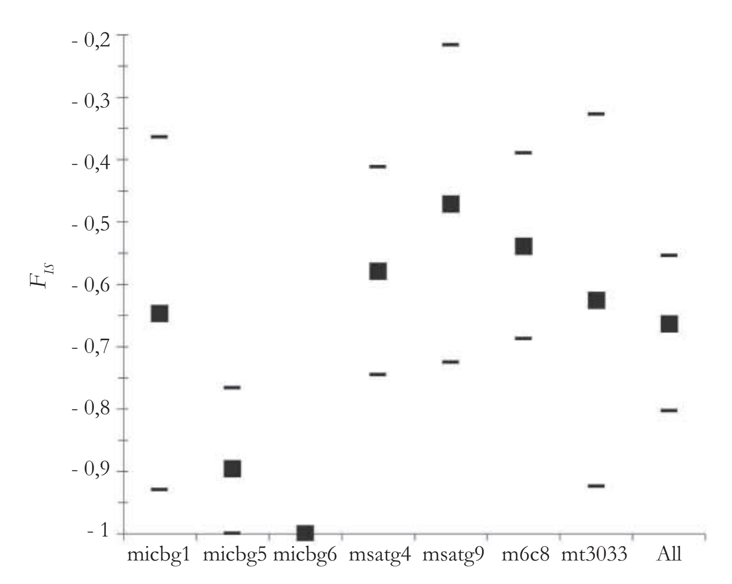


Figure 87
Valeurs de F_{IS} par locus et sur l'ensemble (All) sans le locus trbpa, intervalles de confiance à 95 % de jackknife sur les sous-échantillons (pour les loci) ou de bootstrap sur les loci (pour la moyenne globale : All).

ce qui se passe dans le fichier de sortie "T-BruceiFoyAnCI&Guin-CleanLoci.out", on obtient la figure 87. On voit que le F_{IS} = - 0,66 avec un intervalle de bootstrap à 95 % de [- 0,8, - 0,55]. C'est plus bas que Koffi *et al.*, mais parce que nous avons gardé micbg6. La variance reste apparemment importante entre loci. Cette forte variance entre loci et d'un sous-échantillon à l'autre pourrait être le signe d'événements rares de sexe dans un système très majoritairement clonal, comme le montrent les simulations de Balloux *et al.* (2003).

Cela pourrait provenir également d'allèles nuls rares (il y a quelques rares homozygotes). Ce pourrait être aussi la conséquence d'un taux de mutation variable entre loci. En effet, chez les clones purs, il existe une relation directe entre diversité génétique et F_{IS}. Reprenons la formule générale du F_{IS} :

$$F_{IS} = \frac{Q_I - Q_S}{1 - Q_S}$$

Or nous savons que chez les clones purs, l'homozygotie Q_I tend vers 0, ce qui donne :

$F_{IS} = \frac{-Q_S}{1-Q_S}$, et comme Q_S = 1 - H_s on a forcément $F_{IS} = \frac{-1+H_s}{1-1+H_s} = \frac{-1+H_s}{H_s}$

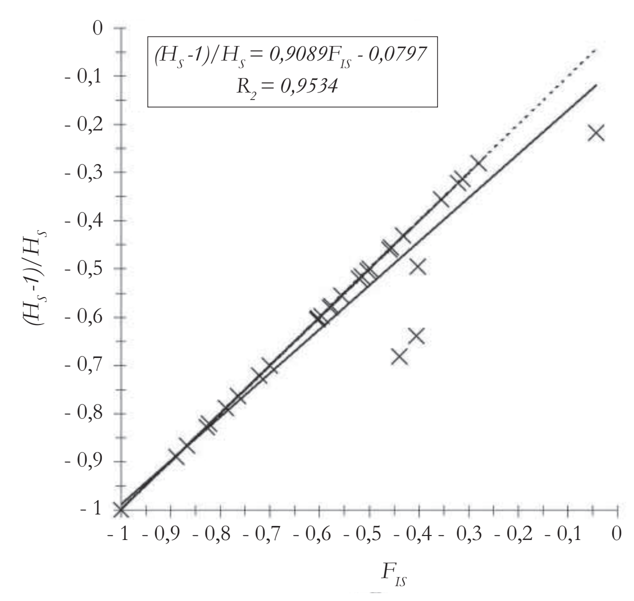


Figure 88
Résultat de la régression entre les valeurs de F_{IS} aux différents loci et dans les différents sous-échantillons et la valeur attendue en fonction de H_s sous l'hypothèse d'une clonalité absolue. La droite d'ajustement parfait est en pointillé.

Dans la figure 88, on remarque une relation quasi parfaite entre les deux paramètres, hormis quatre apostats (cherchez dans le dictionnaire !) dus à quelques individus homozygotes (un pour msatg9, quatre pour m6c8) rencontrés ça et là et très vraisemblablement dus à des *allelic dropouts*, ou à de l'homoplasie (homozygoties fortuites dues au nombre limité d'allèles). Tous les autres points sont en effet parfaitement alignés sur la droite d'ajustement parfait.

La clonalité pure est donc ici certaine.

DIFFÉRENCIATION GÉNÉTIQUE ET STRUCTURE DES POPULATIONS

En général, je préconise les approches globales plutôt que par paire de sous-échantillons. En effet, les mesures et tests par paire ne sont pas les plus efficaces pour appréhender la structure d'une population. Il vaut mieux alors utiliser des distances génétiques. Cependant ici, il n'y a que trois sous-populations subdivisées, parfois en

Tableau 37
Résultats des mesures de différenciation par paire d'échantillons de Tbg1 aux échelles spatiales et temporelles. Tous les tests restant significatifs au seuil Bonferroni séquentiel (en considérant sept tests) sont indiqués en gras. La mesure standardisée du F_{ST}, $F_{ST}' = F_{ST}/(1 - H_s)$ est aussi indiquée.

Échelle	Marqueur	Géographie	Sous-échantillon		F_{ST}	*P*-value	H_s	F_{ST}'
Temporelle	Loci	Bonon	2000	2002	0,0096	0,0182	0,5959	0,0238
		Bonon	2000	2004	0,0160	**0,0063**	0,6129	0,0413
		Bonon	2002	2004	0,0031	0,1836	0,6119	0,0080
		Dubréka	1998	2002	0,0352	0,0330	0,6594	0,1033
	MLG	Bonon	2000	2002	0,1157	**0,0010**	0,8418	0,7311
		Bonon	2000	2004	0,1140	**0,0009**	0,8592	0,8094
		Bonon	2002	2004	0,0250	0,0590	0,9380	0,4032
		Dubréka	1998	2002	0,1006	**0,0059**	0,8570	0,7033
Spatiale	Loci	Entre pays	Bonon	Boffa	0,2940	**0,0001**	0,5760	0,6934
2002			Bonon	Dubréka	0,2127	**0,0001**	0,6177	0,5564
		Guinée	Boffa	Dubréka	0,0514	**0,0017**	0,5988	0,1281
	MLG	Entre pays	Bonon	Boffa	0,1769	**0,0001**	0,8783	1,0000
			Bonon	Dubréka	0,1207	**0,0153**	0,9219	1,0000
		Guinée	Boffa	Dubréka	0,0452	**0,0203**	0,8795	0,3751

deux ou trois périodes (années) d'échantillonnage. J'ai donc créé un nouveau fichier contenant les sept loci ne présentant pas de problème et les MLG ("T-BruceiFoyer AnCleanLoci&MLGCreate.txt"). Il faut mettre ces données au format Fstat et analyser les différenciations par paire de sous-échantillons en sélectionnant les loci de façon pertinente (ne pas laisser MLG avec les loci normaux !). En fait, les données "loci" sont déjà dans "T-BruceiFoyAnCI&Guin-CleanLoci.dat". Pour les MLG, il suffit d'ouvrir le fichier global et sélectionner le locus MLG avec le menu "Options" et "Loci to use". Pour ces deux nouveaux fichiers, l'analyse se fait sous Fstat avec la procédure "Pairwise test of differentiation" avec 10 000 permutations des individus entre sous-populations et les "Fst per pair of samples". Le résultat de ces deux analyses, si vous ne vous êtes pas trompés, à extraire des fichiers *.fst et *.pvl, sont compilés dans le tableau 37.

La différenciation temporelle est plus facile a détecter avec les MLGs. Substantielle au bout de deux années, elle devient très importante au bout de quatre ans. La dérive

est donc rapide et suggère de faibles effectifs efficaces pour les MLGs. La structure géographique est très prononcée avec un isolement total entre Guinée et Côte d'Ivoire et probablement peu d'échanges entre Boffa et Dubréka. Si on reprend l'équation (26) du chapitre 2 de la première partie (modèle en deux îles), on peut en déduire un équivalent $Nm = (1 - F_{ST}')/8F_{ST}' = 0{,}21$ MLG échangé par génération entre deux sous-populations. Il est probable que la division cellulaire n'est pas la bonne mesure du temps de générations ici. En effet, cela signifierait que chaque cellule de trypanosome correspond à un individu. Étant donné le nombre de personnes atteintes, et surtout le nombre de cellules trypanosomiales contenues par patient, cela reviendrait à des populations de tailles gigantesques qui ne devraient pas ou peu dériver (pour des chiffres, consulter l'article original de Koffi *et al.*, 2009). Le temps de génération correspond donc davantage au temps d'un cycle complet tsé-tsé-homme-tsé-tsé qui prend environ 37 à 49 jours (se référer à l'article de Koffi *et al.*, 2009 pour les détails), d'où un nombre maximal de générations par an de 10. Soit donc deux MLG échangés par année.

Calculs d'effectifs efficaces

Nous allons utiliser ici une pirouette dont nous vérifierons la pertinence ensuite à l'aide de quelques simulations. Comme nous avons des échantillons des mêmes foyers échantillonnés dans le temps pour Bonon et Dubréka, nous allons tenter d'estimer la taille de dérive des MLGs à l'aide de méthodes temporelles et spatio/temporelles. Pour Bonon et pour Dubréka, nous utiliserons la méthode de Waples (1989) avec NeEstimator. Pour les foyers guinéens, nous pourrons aussi essayer d'estimer conjointement la taille efficace et le taux de migration à l'aide de la méthode de Wang et Whitlock (2003) avec le logiciel MLNe. Il nous faut donc dans un premier temps convertir les données MLG au format approprié.

Construction des fichiers pour NeEstimator et pour MLNe

Pour la méthode de Waples (NeEstimator), il faut faire un fichier de type Genepop pour chaque année de chaque site pertinent, soit cinq fichiers (Bonon en 2000, 2002, 2004, Dubréka 1998 et Dubréka 2002), comme en figure 89.

Pour MLNe nous allons passer par Create, car le formatage du fichier est horrible (je ne remercierai jamais assez Jason Coombs[16]). Le fichier a donc la forme de la figure 90.

Il convient ensuite sous Create de charger ce fichier et de lui donner les informations, comme indiqué dans la figure 91.

Create vous demande si c'est bon en vous montrant ce qu'il a fait et vous dites oui. Un nouveau cadre apparaît où vous allez cocher "MLNE" dans "Specialized gene-

[16] Notez que comme PGD-Spider ne prend pas en charge cette conversion, CREATE est donc à ma connaissance le seul logiciel utilisable pour convertir un jeu de données au format MLNe.

```
MLG¶
Bon00»      ,»        014014¶
Bon00»      ,»        021021¶
Bon00»      ,»        022022¶
Bon00»      ,»        022022¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        026026¶
Bon00»      ,»        027027¶
Bon00»      ,»        034034¶
Bon00»      ,»        036036¶
Bon00»      ,»        040040¶
Bon00»      ,»        042042
```

Figure 89
Aspect d'un fichier de données pour NeEstimator. Exemple des données de Bonon 2000. Le seul locus correspond aux MLGs. Les données commencent en seconde ligne (qui ne sera pas lue par NeEstimator). Le signe ">>" signifie une tabulation.

tic programs" et cliquer ensuite sur "Create". On vous demande ensuite de choisir les populations focales (pour laquelle le N_e et le m seront calculés) et sources (d'immigrants). Nous choisissons d'abord Bonon comme population focale (pour laquelle nous essayerons d'obtenir m et N_e) et les deux autres comme source (fig. 92).

Renommez le fichier de telle sorte qu'il soit identifié comme focalisé sur Bonon, comme par exemple "T-BruceiFoyerAnMLGCreate-MLNE-Bonon.txt". Faites ensuite la même chose pour Boffa et Dubréka. Pour Boffa ça ne marche pas, car il n'y a qu'un seul échantillon temporel. Nous n'obtenons donc que deux fichiers analysables par MLNe, un pour Bonon et un pour Dubréka. N'oubliez pas d'identifier le fichier de Dubréka.

Analyses avec NeEstimator

Lancez NeEstimator et après avoir lu l'avertissement, cliquez sur OK. Après avoir cliqué sur "File" et choisi "New", vous obtenez un cadre de menu où vous allez sélectionner les mêmes options que celles indiquées en figure 93. En particulier, choisissez le format de fichier Genepop et ignorez la première ligne avec un format de délimitation entre données "Tab" (tabulations).

Cliquez ensuite sur l'onglet "Data file" puis sur "Load". Allez chercher les fichiers contenant les données de Bonon 2000 auxquelles vous affecterez la génération 0 et Bonon 2002 auxquelles vous affecterez la génération 19 (10 générations par an, comme indiqué plus haut), comme représenté dans la figure 94.

Pop»	t»	MLG»	MLG¶
Bon»	39»	014»	014¶
Bon»	39»	021»	021¶
Bon»	39»	022»	022¶
Bon»	39»	022»	022¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	026»	026¶
Bon»	39»	027»	027¶
Bon»	39»	034»	034¶
Bon»	39»	036»	036¶
Bon»	39»	040»	040¶
Bon»	39»	042»	042¶
Bon»	59»	008»	008¶
Bon»	59»	012»	012¶
Bon»	59»	015»	015¶
Bon»	59»	017»	017¶
Bon»	59»	020»	020¶
Bon»	59»	022»	022¶
Bon»	59»	029»	029¶
Bon»	59»	008»	008¶
Bon»	59»	010»	010¶
Bon»	59»	010»	010¶
Bon»	59»	022»	022¶
Bon»	59»	023»	023¶
Bon»	59»	029»	029¶
Bon»	59»	045»	045¶
Bon»	79»	001»	001¶
Bon»	79»	002»	002¶
Bon»	79»	004»	004¶
Bon»	79»	005»	005¶
Bon»	79»	007»	007¶
Bon»	79»	015»	015¶
Bon»	79»	024»	024¶
Bon»	79»	032»	032¶
Bon»	79»	035»	035¶
[illegible]	[illegible]	[illegible]	[illegible]

Figure 90
Le jeu de données MLG de tous les sous-échantillons pour Create, avant transformation pour MLNe. t indique la génération en partant de 0 pour 1998 et en finissant avec 79 pour 2004 sur la base de 10 générations par an.

Il s'agit ensuite de lancer le calcul en cliquant sur "File" et "Run", comme sur la figure 95.

Les résultats apparaissent sous forme d'un tableau (fig. 96). Seule l'analyse par la méthode temporelle de Waples (celle qui nous intéresse ici) donne un résultat avec 95 % d'intervalle de confiance. Cet intervalle de confiance est calculé selon la formulation complexe décrite dans Waples (1989) qui utilise la loi du Chi-2 avec un degré de liberté égal au nombre total d'allèles indépendants ayant servi à l'estimation et un seuil α = 0,05.



Figure 91
Menu Create pour créer le fichier pour MLNe.

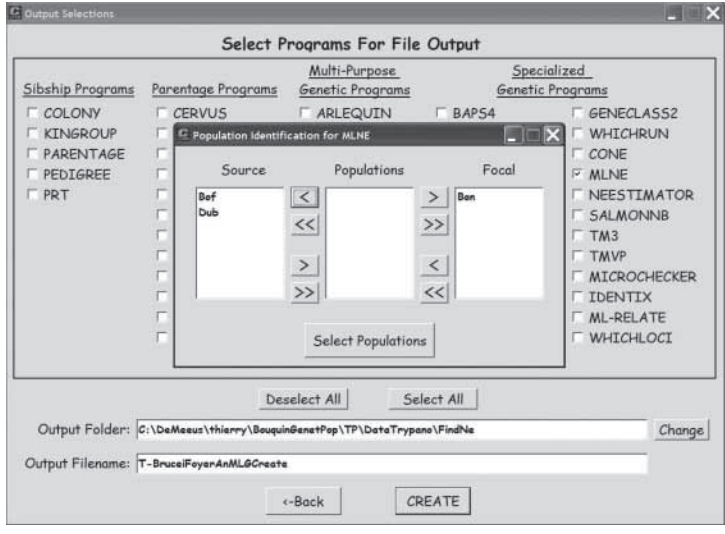


Figure 92
Définir la population focale et les populations sources pour MLNe dans CREATE.

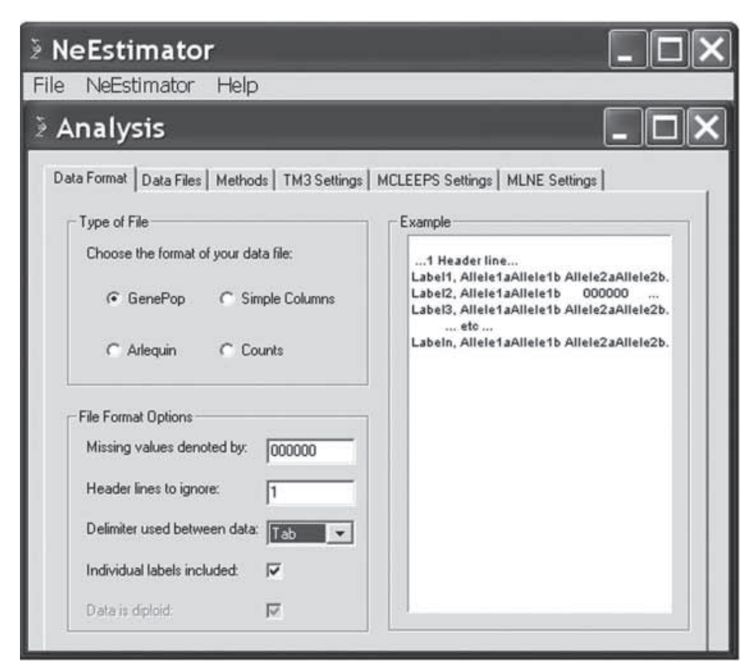


Figure 93
Menu NeEstimator pour estimation de N_e temporel (Waples).

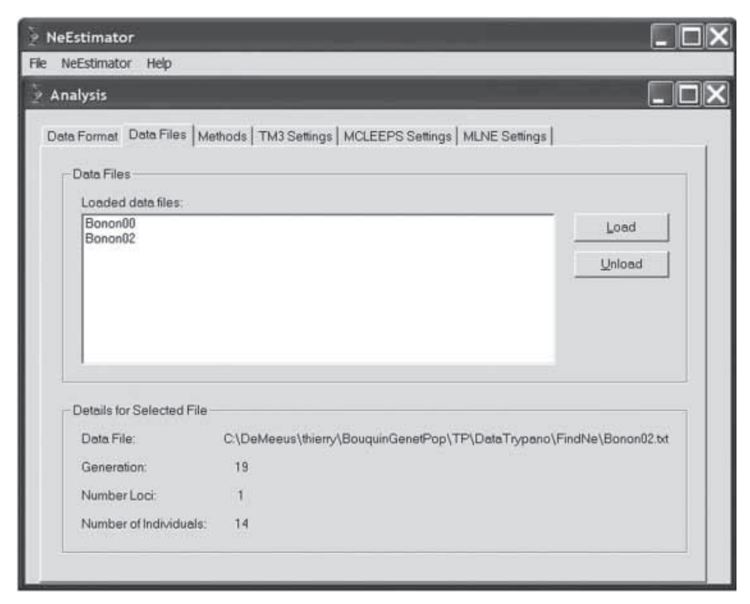


Figure 94
Cadre de menu de NeEstimator pour choisir les fichiers à analyser pour un calcul d'effectifs efficaces pour deux échantillons du même site prélevés à deux dates différentes.

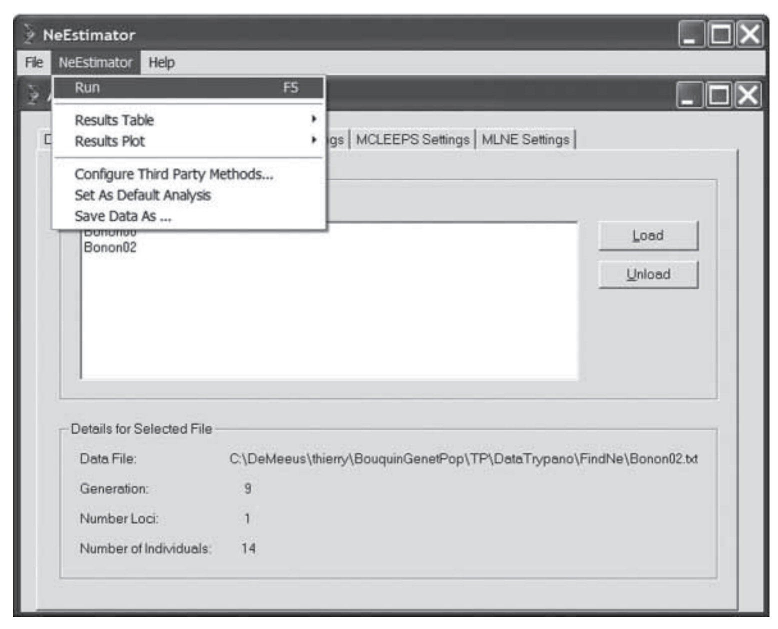
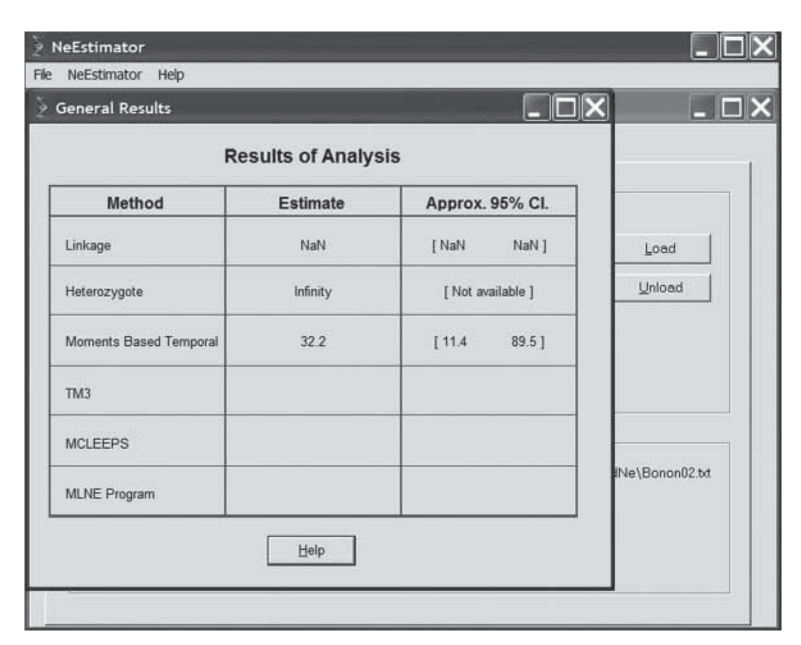


Figure 95
Lancement du calcul de N_e.



Method	Estimate	Approx. 95% CI.
Linkage	NaN	[NaN NaN]
Heterozygote	Infinity	[Not available]
Moments Based Temporal	32.2	[11.4 89.5]
TM3		
MCLEEPS		
MLNE Program		

Figure 96
Résultats de l'analyse NeEstimator pour le calcul du N_e temporel de Waples à Bonon.

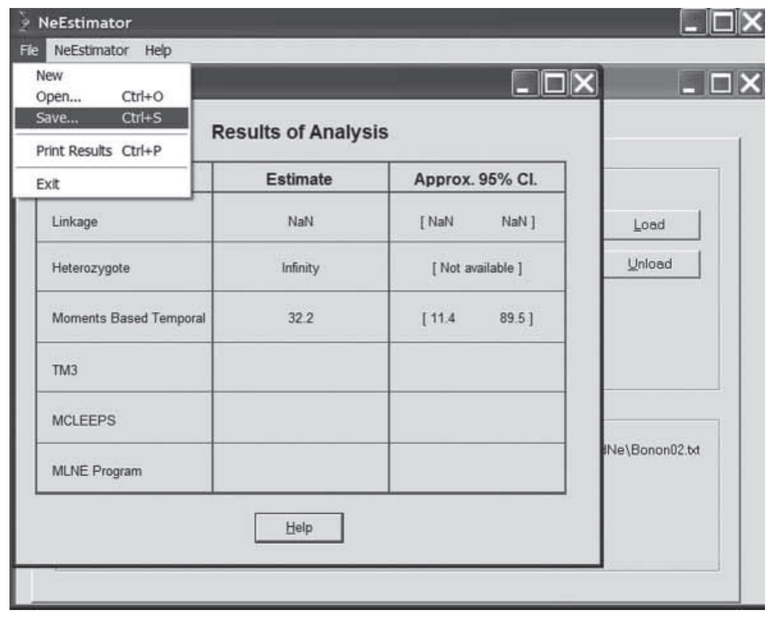


Figure 97
Sauver les résultats de NeEstimator.

Vous pouvez (et je le conseille) sauvegarder ces résultats avec le menu déroulant “File” et “Save” (fig. 97). Nommez votre fichier de façon appropriée et NeEstimator y ajoutera l’extension NeA. J’ai personnellement nommé ce fichier “ResNeEstimBonon00-02.NeA”.

Refaites la même chose pour tous les sous-échantillons temporels. Les résultats sont synthétisés dans le tableau 38.

Analyses avec MLNE

Après avoir créé un répertoire pour Bonon et pour Dubréka et y avoir déplacé les fichiers correspondants créés par Create, copiez dans chacun de ces deux répertoires le logiciel MLNE “mne2.exe”. Lisez bien la notice, qui n’est pas des plus didactiques, afin d’effectuer les modifications nécessaires dans les fichiers sources. Prenez le fichier pour Bonon. La première ligne doit indiquer “1”, car vous souhaitez estimer à la fois m et N_e. La deuxième ligne indique la taille efficace maximale autorisée (pour économiser de la mémoire), et est par défaut 5 000, ce qui est largement suffisant. Si le résultat est proche de cette valeur, vous pourrez éventuellement recommencer avec une valeur plus élevée. La troisième ligne n’a pas d’intérêt et on ne s’en occupe pas. La quatrième ligne est destinée aux informaticiens experts dont nous ne faisons malheureusement pas partie, donc nous zappons. La cinquième ligne désigne le nombre de loci (vérifiez que le nombre indiqué est bien “1”). La sixième ligne

indique le nombre total d'allèles. La septième ligne indique le nombre de sous-échantillons temporels pour la sous-population focale (ici Bonon). Il y en a trois correspondants aux générations 39, 59 et 79. Il faut donc que soit indiqué "3". En huitième ligne sont indiqués les numéros de cohorte de chacun de ces sous-échantillons temporels, dans l'ordre et en commençant par "0". Il faut donc taper "0,20,40" sur cette ligne. Ensuite, ce sont les données codées par Create au format MLNE et que personnellement je n'aurais jamais eu le courage de faire tout seul. Il faut ensuite enregistrer ce fichier sous le nom "MNE_DATA" en lettres capitales et sans extension. Il suffit ensuite de double cliquer sur mne2.exe pour lancer la procédure. Après un certain nombre de calculs plus ou moins longs, le logiciel crée alors un fichier "MNE_OUT". Le programme donne les valeurs de N_e et de m selon deux méthodes. Celle du maximum de vraisemblance avec les intervalles de confiance à 95 % et celle des moments. Ces deux méthodes sont décrites dans l'article Wang et Whitlock (2003). Pour Dubréka, il n'y a que deux sous-échantillons temporels ("2" en ligne 7) correspondant aux cohortes 0 et 59 ("0,59" en ligne 8). Les résultats de cette approche figurent dans le tableau 38.

Estimation de la taille clonale des foyers par modélisation

Ici, les allergiques aux formules mathématiques vont souffrir, mais il n'y a guère d'autres moyens d'expliquer comment obtenir des valeurs d'effectifs clonaux. Ceux pour lesquels la cause est perdue peuvent se référer directement aux résultats finaux. Cependant, si vous lisez ce chapitre c'est que vous comptez travailler sur des organismes à reproduction clonale. Je crois alors indispensable d'avoir compris au moins une fois ce qui suit, ou au moins de comprendre la démarche permettant d'aboutir aux résultats finaux.

Cas général

Dans un modèle en île subdivisé en n sous-populations, chacune composée de N individus diploïdes à générations non chevauchantes avec un taux de mutation u dans un modèle IAM (*infinite allele model*), soit Q_I la probabilité de prendre au hasard deux fois le même allèle au sein d'un même individu, Q_S la probabilité de prélever au hasard le même allèle dans deux individus de la même sous-population et Q_T la probabilité de prendre deux allèles identiques dans deux sous-populations différentes de la population totale, soit $\gamma = (1 - u)^2$ la probabilité qu'aucun des deux allèles pris au hasard n'ait muté d'une génération à l'autre, c la proportion de zygotes formés de façon clonale (asexuée) et s la proportion, parmi les $(1 - c)$ qui se forment suite à une autofécondation, soit q_s la probabilité de tirer au hasard deux individus de la même sous-population qui soient originaires tous les deux d'une seule et même sous-population avant migration et q_d la probabilité que deux individus pris au hasard dans deux sous-populations différentes parmi les n disponibles soient issus,

avant migration, de la même sous-population, alors la récurrence d'une génération à l'autre pour Q_I, Q_S et Q_T peut s'écrire :

$$\begin{cases} Q_{I(t+1)} = \gamma\left\{cQ_{I(t)} + (1-c)\left[s\left(\dfrac{1+Q_{I(t)}}{2}\right) + (1-s)Q_{S(t)}\right]\right\} \\ Q_{S(t+1)} = \gamma\left\{q_s\left[\dfrac{1}{N}\left(\dfrac{1+Q_{I(t)}}{2}\right) + \left(1-\dfrac{1}{N}\right)Q_{S(t)}\right] + (1-q_s)Q_{T(t)}\right\} \\ Q_{T(t+1)} = \gamma\left\{q_d\left[\dfrac{1}{N}\left(\dfrac{1+Q_{I(t)}}{2}\right) + \left(1-\dfrac{1}{N}\right)Q_{S(t)}\right] + (1-q_d)Q_{T(t)}\right\} \end{cases} \quad (69)$$

Pour que deux allèles restent identiques, il faut qu'aucun des deux n'ait muté (nous négligeons l'homoplasie), soit γ. Pour Q_I, les zygotes issus de reproduction clonale (probabilité c) gardent la même probabilité de posséder deux allèles identiques qu'à la génération précédente. Parmi ceux issus de reproduction sexuée (1 - c), ceux issus d'autofécondations (s) ont déjà deux gènes identiques qui le restent avec la probabilité $Q_{I(t)}$ ou, sachant qu'ils n'étaient pas identiques (1 - $Q_{I(t)}$), la probabilité de tirer deux fois le même après autofécondation est de ½, soit donc $Q_{I(t)}$ + (1 - $Q_{I(t)}$)/2 = (1 + $Q_{I(t)}$)/2. Les zygotes issus de croisements panmictiques (1 - s) obtiennent deux allèles identiques avec la probabilité $Q_{S(t)}$, par définition. Pour Q_S, la probabilité de tirer deux allèles identiques de deux individus de la même sous-population, il faut que ces deux individus aient été issus de la même sous-population (q_s). Parmi ceux-ci, on tire deux fois le même individu (1/N) et ce dernier a les deux même allèles avec la probabilité $Q_{I(t)}$ ou ils ne le sont pas (1 - $Q_{I(t)}$) et on tire deux fois le même avec la probabilité ½, ce qui donne (1/N) (1 + $Q_{I(t)}$)/2, mais si on tire deux individus différents (1 - 1/N), la probabilité de tirer deux allèles identiques est $Q_{S(t)}$ par définition, ce qui donne bien (1 - 1/N)$Q_{S(t)}$ et donc au final, si deux individus sont issus d'une même sous-populations (q_s), la probabilité de tirer deux allèles identiques chez eux est (1/N)(1 + $Q_{I(t)}$)/2 + (1 - 1/N)$Q_{S(t)}$. Enfin, si les deux individus n'étaient pas initialement dans la même sous-population (1-q_s), alors la probabilité de tirer deux fois le même allèle est $Q_{T(t)}$ par définition. Pour finir, en ce qui concerne Q_T, les deux individus tirés de deux sous-populations différentes pouvaient initialement avoir été dans la même sous-population (q_d) et dans ce cas, la probabilité de tirer deux allèles identiques chez eux est (1/N)(1 + $Q_{I(t)}$)/2 + (1 - 1/N)$Q_{S(t)}$ (comme vu précédemment), alors que dans le cas contraire (1-q_d), la probabilité de tirer deux allèles identiques est $Q_{T(t)}$ par définition.

Nous allons supposer que les sous-populations sont de taille N suffisamment grande de telle sorte qu'échantillonner dans une telle sous-population ne change pas les fréquences d'allèles. Nous savons qu'il y a n sous-populations. Alors, q_s représente la probabilité que soit les deux individus pris au hasard dans une sous-population soient tous les deux non migrants, avec la probabilité $P_1 = (1 - m)(1 - m) = (1 - m)^2$ et auquel cas ils sont effectivement issus de la même sous-population avant migration, soit que ces

deux individus soient des immigrants venus d'autres sous-populations, avec une probabilité $P_2 = m^2$ et qu'ils viennent d'une même sous-population parmi les $(n - 1)$ restantes, soit $P_3 = 1/(n - 1)^2$, mais sachant que les $(n - 1)$ sous-populations peuvent indépendamment fournir ces deux individus. Par conséquent, $q_s = P_1 + P_2 \times P_3 \times (n - 1)$, ou :

$$q_s = (1-m)^2 + \frac{m^2}{(n-1)} \tag{70}$$

Par ailleurs, q_d est égal à la probabilité de prélever deux individus de deux sous-populations différentes parmi les n puis parmi les $(n - 1)$ disponibles avec n possibilités, soit $P_4 = (1/n)(1/(n - 1)) \times n = 1/(n - 1)$ et que tous les deux soient des migrants (m^2) et que, avant migration, l'un provienne alors d'une des $n - 1$ sous-populations restantes et l'autre de cette même sous-population parmi les n-2 restantes (soit $\frac{1}{n-1}(n-1)\frac{1}{n-1}(n-2)$), soit $P_5 = m^2 \frac{1}{(n-1)^2}(n-1)(n-2) = \frac{m^2(n-2)}{n-1}$ ou bien alors que le premier individu soit un immigrant et pas l'autre ou l'inverse $(2m(1 - m))$ et que l'immigrant provienne d'une autre des $(n - 1)$ sous-populations $(1/(n - 1))$ avec $n - 1$ possibilités, donc $P_6 = 2m(1 - m)(n - 1)/(n - 1) = 2m(1 - m)$. Par conséquent, nous pouvons écrire que $q_d = P_4 \times (P_5 + P_6)$, ou encore :

$$q_d = \frac{1}{n-1}\left[\frac{m^2(n-2)}{n-1} + 2m(1-m)\right]$$

Nous pouvons réarranger cette équation :

$$q_d = \frac{1}{n-1}\left[2m(1-m) + m^2\frac{n-1-1}{n-1}\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[2m(1-m) + m^2\left(1 - \frac{1}{n-1}\right)\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[2m(1-m) + m^2 - \frac{m^2}{n-1}\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[2m - 2m^2 + m^2 - \frac{m^2}{n-1}\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[2m - m^2 - \frac{m^2}{n-1}\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[1 - 1 + 2m - m^2 - \frac{m^2}{n-1}\right]$$

$$\Leftrightarrow$$

$$q_d = \frac{1}{n-1}\left[1 - (1-m)^2 - \frac{m^2}{n-1}\right]$$

Il en résulte que :

$$q_d = \frac{1 - q_s}{n-1} \tag{71}$$

Nous faisons maintenant l'hypothèse d'une clonalité totale (c = 1), les récurrences deviennent :

$$\begin{cases} Q_{I(t+1)} = \gamma c Q_{I(t)} \\ Q_{S(t+1)} = \gamma \left\{ q_s \left[\frac{1}{N}\left(\frac{1+Q_{I(t)}}{2}\right) + \left(1 - \frac{1}{N}\right) Q_{S(t)} \right] + (1-q_s)\, Q_{T(t)} \right\} \\ Q_{T(t+1)} = \gamma \left\{ q_d \left[\frac{1}{N}\left(\frac{1+Q_{I(t)}}{2}\right) + \left(1 - \frac{1}{N}\right) Q_{S(t)} \right] + (1-q_d)\, Q_{T(t)} \right\} \end{cases} \tag{72}$$

Si nous nous posons à un état proche de l'équilibre mutation/migration/dérive, alors $Q_{I(t)} = Q_{I(t+1)} = Q_I$; $Q_{S(t)} = Q_{S(t+1)} = Q_S$; $Q_{T(t)} = Q_{T(t+1)} = Q_T$ et on voit tout de suite que $Q_I = 0$, ce qui correspond bien à l'attendu théorique d'une population clonale en nombre infini d'allèles (hétérozygotie totale) (Balloux *et al.*, 2003). Le système d'équations précédent devient :

$$\begin{cases} Q_I = 0 \\ Q_S = \gamma \left\{ q_s \left[\frac{1}{2N} + \left(1 - \frac{1}{N}\right) Q_S \right] + (1 - q_s) Q_T \right\} \\ Q_T = \gamma \left\{ q_d \left[\frac{1}{2N} + \left(1 - \frac{1}{N}\right) Q_S \right] + (1 - q_d) Q_T \right\} \end{cases} \tag{73}$$

On peut résoudre ce système de deux équations à deux inconnues à l'aide des calculs matriciels comme dans l'article de Balloux *et al.* (2003). Cependant, pour gagner du temps et simplifier les calculs nous allons tout de suite faire trois hypothèses (trois cas de figures) et voir ce que cela donne. Dans la première hypothèse, nous supposerons que le nombre de sous-populations n est très grand. Dans le deuxième cas

qu'il n'y a que deux sous-populations, telles que Boffa et Dubréka en Guinée (il existe un troisième foyer, Forecariah, mais qui reste assez éloigné) et en Côte d'Ivoire avec Bonon et Sinfra (Kaba *et al.*, 2006). Enfin, dans la mesure où nous avons pu constater que la différenciation entre foyers était assez forte nous ferons, pour le troisième cas de figure, l'hypothèse d'une seule population isolée.

Nombre infini de sous-populations

C'est le modèle décrit dans De Meeûs et Balloux (2005). Dans ce cas, on montre que, puisque $n \to \infty$:

$$q_s = (1-m)^2 + \frac{m^2}{(n-1)} \approx (1-m)^2$$

et

$$q_d = \frac{1-q_s}{n-1} \approx 0$$

Le système de trois équations (73) devient :

$$\begin{cases} Q_I = 0 \\ Q_S = \gamma\left\{(1-m)^2\left[\frac{1}{2N} + \left(1-\frac{1}{N}\right).Q_S\right] + \left[1-(1-m)^2\right].Q_T\right\} \\ Q_T = \gamma Q_T \end{cases}$$

Il y apparaît clairement que la solution pour Q_T est $Q_T = 0$ et donc :

$$Q_S = \gamma\left\{(1-m)^2\left[\frac{1}{2N} + \left(1-\frac{1}{N}\right)Q_S\right]\right\}$$

À partir de là on peut poser :

$$Q_S\left[1-\gamma(1-m)^2\left(1-\frac{1}{N}\right)\right] = \gamma(1-m)^2\frac{1}{2N}$$

d'où il est facile d'extraire :

$$Q_S = \frac{\gamma(1-m)^2\frac{1}{2N}}{1-\gamma(1-m)^2\left(1-\frac{1}{N}\right)}$$

Nous pouvons réarranger cette équation :

$$Q_S = \frac{\frac{\gamma(1-m)^2}{2N}}{\frac{2N-\gamma(1-m)^2(2N-2)}{2N}} = \frac{\gamma(1-m)^2}{2N-\gamma(1-m)^2(2N-2)}$$

Sachant que $\gamma = (1 - u)^2$, nous pouvons poser :

$$Q_S = \frac{(1-u)^2(1-m)^2}{2N - 2N(1-u)^2(1-m)^2 + 2(1-u)^2(1-m)^2}$$

Nous allons considérer maintenant que tous les termes en u^2 et m^2 sont négligeables devant 1. L'équation précédente peut donc s'écrire :

$$Q_S = \frac{(1-2u)(1-2m)}{2N - 2N(1-2u)(1-2m) + 2(1-2u)(1-2m)}$$

En développant nous obtenons :

$$Q_S = \frac{(1-2m-2u+4um)}{2N - 2N(1-2m-2u+4um) + 2(1-2m-2u+4um)}$$

Nous pouvons également négliger les termes en *um* devant 1, ce qui donne :

$$Q_S = \frac{(1-2m-2u)}{2N - 2N(1-2m-2u) + 2(1-2m-2u)}$$

$$Q_S = \frac{(1-2m-2u)}{2N - 2N + 4N(m+u) + 2(1-2m-2u)}$$

$$Q_S = \frac{(1-2m-2u)}{4N(m+u) + 2(1-2m-2u)}$$

Nous allons maintenant considérer que le taux de migration est faible (c'est le cas ici) et le taux de mutation aussi. Le taux de mutation moyen des microsatellites est en effet de l'ordre de $u = 10^{-3}$ d'après la littérature sur cette question (Ellegren, 2000 ; Balloux et Lugon-Moulin, 2002 ; Ellegren, 2004). Si nous négligeons les termes en *u* et *m* devant 1 nous obtenons pour Q_S :

$$Q_S \approx \frac{1}{4N(m+u)+2} \qquad (74)$$

Nous pouvons maintenant nous servir de ces valeurs d'identité à l'équilibre pour calculer les F_{IS} et F_{ST} à l'équilibre mutation, migration et dérive en utilisant l'équation (21) du chapitre 2 de la première partie de ce manuel :

$$\begin{cases} F_{IS} = \dfrac{Q_I - Q_S}{1 - Q_S} \approx \dfrac{0 - \dfrac{1}{4N(m+u)+2}}{1 - \dfrac{1}{4N(m+u)+2}} \\ \\ F_{ST} = \dfrac{Q_S - Q_T}{1 - Q_T} \approx \dfrac{\dfrac{1}{4N(m+u)+2} - 0}{1 - 0} \end{cases}$$

ce qui donne :

$$\begin{cases} F_{IS} \approx \dfrac{-\dfrac{1}{4N(m+u)+2}}{\dfrac{4N(m+u)+2-1}{4N(m+u)+2}} \\ F_{ST} \approx \dfrac{1}{4N(m+u)+2} \end{cases}$$

Et finalement

$$\begin{cases} F_{IS} \approx \dfrac{-1}{4N(m+u)+1} \\ F_{ST} \approx \dfrac{1}{4N(m+u)+2} \end{cases} \qquad (75)$$

Nous retrouvons la fameuse équation $F_{ST} = -F_{IS}/(1-F_{IS})$. À partir de là, nous pouvons extraire N et m des valeurs de F_{IS} et F_{ST}.

$$\begin{cases} 4N(m+u)F_{IS} + F_{IS} = -1 \\ 4N(m+u)F_{ST} + 2F_{ST} = 1 \end{cases}$$

<=>

$$\begin{cases} 4N(m+u)F_{IS} = -1 - F_{IS} \\ 4N(m+u)F_{ST} = 1 - 2F_{ST} \end{cases}$$

Nous posons que $m >> u$ et donc :

$$\begin{cases} 4NmF_{IS} \approx -1 - F_{IS} \\ 4NmF_{ST} \approx 1 - 2F_{ST} \end{cases}$$

Nous nous retrouvons donc avec deux valeurs pour Nm :

$$\begin{cases} Nm \approx \dfrac{-1-F_{IS}}{4F_{IS}} \\ Nm \approx \dfrac{1-2F_{ST}}{4F_{ST}} \end{cases} \qquad (76)$$

Nous savons, d'après les simulations de De Meeûs et Balloux (2005), que c'est le F_{IS} qui donne les meilleurs résultats, c'est donc cette formulation que nous retiendrons. Les résultats du calcul des Nm figurent dans le tableau 38. Ils ont nécessité le calcul d'un F_{IS} avec son intervalle de confiance à 95 % de bootstrap dans chaque foyer (valeurs moyennes calculées sur l'ensemble des sous-échantillons). Pour Boffa (un seul sous-échantillon), si on utilise Fstat qui ne sait pas

travailler sur un seul sous-échantillon, il faut ajouter un deuxième sous-échantillon fictif de même taille que Boffa et fixé (111111) pour tous les loci.

Deux sous-populations

Avec seulement deux sous-populations, comme on peut raisonnablement penser que ce soit le cas en Guinée avec Boffa et Dubréka et en Côte d'Ivoire avec Bonon et Sinfra (Kaba *et al.*, 2006), les équations (70), (71) et (73) deviennent :

$$q_s = (1-m)^2 + m^2 = 1 - 2m + 2m^2 = 1 - 2m(1-m) \tag{77}$$

$$q_d = 1 - q_s = 1 - 1 + 2m(1-m) = 2m(1-m) \tag{78}$$

Il n'y a cependant pas de façon simple de résoudre le système d'équations (73) ici et il faut passer par une résolution matricielle avec un logiciel de mathématiques. Ceci avait déjà été fait dans Balloux *et al.* (2003) et donne pour F_{IS} et F_{ST} (après correction des erreurs dans les formules) (Koffi *et al.*, 2009, Appendice) :

$$\begin{cases} F_{IS} = \dfrac{\gamma[q_s - \gamma(q_s - q_d)]}{2N(1-\gamma)[\gamma(q_s - q_d) - 1] - \gamma[q_s - \gamma(q_s - q_d)]} \\[2ex] F_{ST} = \dfrac{\gamma(1-\gamma)(q_s - q_d)}{2N(1-\gamma)[1 - \gamma(q_s - q_d)] + \gamma[q_d(2\gamma - 1) - 2q_s(\gamma - 1)]} \end{cases}$$

Si on remplace q_d par 1 - q_s (dans le cas où n = 2 sous-populations) :

$$\begin{cases} F_{IS} = \dfrac{\gamma[q_s - \gamma(2q_s - 1)]}{2N(1-\gamma)[\gamma(2q_s - 1) - 1] - \gamma[q_s - \gamma(2q_s - 1)]} \\[2ex] F_{ST} = \dfrac{\gamma(1-\gamma)(2q_s - 1)}{2N(1-\gamma)[1 - \gamma(2q_s - 1)] + [(1 - q_s)(2\gamma - 1) - 2q_s(\gamma - 1)]} \end{cases}$$

Sachant que les termes en u^2 sont négligeables par rapport à 1, on peut considérer que $\gamma \approx 1 - 2u$ et donc :

$$\begin{cases} F_{IS} = \dfrac{(1-2u)[q_s - (1-2u)(2q_s - 1)]}{2N(1-1+2u)[(1-2u)(2q_s - 1) - 1] - (1-2u)[q_s - (1-2u)(2q_s - 1)]} \\[2ex] F_{ST} = \dfrac{(1-2u)(1-1+2u)(2q_s - 1)}{2N(1-1+2u)[1 - (1-2u)(2q_s - 1)] + [(1 - q_s)(2 - 2u - 1) + 2q_s(1 - 2u - 1)]} \end{cases}$$



$$\begin{cases} F_{IS} = \dfrac{(1-2u)(q_s - 2q_s + 1 + 4uq_s - 2u)}{4Nu[2q_s - 1 - 4uq_s + 2u - 1] - (1-2u)(q_s - 2q_s + 1 + 4uq_s - 2u)} \\ F_{ST} = \dfrac{(2u - 4u^2)(2q_s - 1)}{4Nu[1 - 2q_s + 1 + 4uq_s - 2u] + [(1-q_s)(1-2u) - 4uq_s]} \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} F_{IS} = \dfrac{(1-2u)(-q_s + 1 + 4uq_s - 2u)}{8Nu[q_s - 1 - 2uq_s + u] - (1-2u)(-q_s + 1 + 4uq_s - 2u)} \\ F_{ST} = \dfrac{4uq_s - 2u - 8u^2 q_s + 4u^2}{8Nu[1 - u - q_s(1-2u)] + [1 - 2u - q_s + 2uq_s - 4uq_s]} \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} F_{IS} = \dfrac{-q_s + 1 + 4uq_s - 2u + 2uq_s - 2u - 8u^2 q_s + 4u^2}{8Nu[-(1-u) + q_s(1-2u)] - (1-2u)(1 - 2u - q_s(1-4u))} \\ F_{ST} = \dfrac{2u[2q_s(1-2u) - (1-2u)]}{8Nu[1 - u - q_s(1-2u)] + [(1-2u) - q_s(1+2u)]} \end{cases}$$

Nous allons maintenant négliger les termes en u devant 1, ce qui donne :

$$\begin{cases} F_{IS} = -\dfrac{1-q_s}{8Nu(1-q_s) + (1-q_s)} \\ F_{ST} = \dfrac{2u[2q_s - 1]}{8Nu(1-q_s) + (1-q_s)} \end{cases} \tag{79}$$

Nous allons maintenant poser que $q_s = 1 - 2m(1 - m) < 1$. Cette valeur maximale correspond ici à $m < 1$. En effet, dans le cas de deux sous-populations, $m = 1$ est équivalent à une absence de migration puisque cela signifie que tous les individus d'une sous-population migrent ensemble dans l'autre et vice-versa. Nous allons en fait ne considérer que les cas où $0 \le m \le 0{,}5$, où $m = 0{,}5$ correspond dans ce cas au maximum d'échange de migrants possible. Nous pouvons donc simplifier le système d'équations (79) en :

$$\begin{cases} F_{IS} = -\dfrac{1}{8Nu+1} \\ F_{ST} = \dfrac{2u[2q_s - 1]}{8Nu(1-q_s) + (1-q_s)} \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} 8NuF_{IS} + F_{IS} = -1 \\ (8Nu+1)(1-q_s)F_{ST} - 4uq_s = -2u \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} N = -\dfrac{F_{IS}+1}{8uF_{IS}} \\ q_s\left[-(8Nu+1)F_{ST} - 4u\right] = -2u - (8Nu+1)F_{ST} \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} N = -\dfrac{F_{IS}+1}{8uF_{IS}} \\ q_s = \dfrac{(8Nu+1)F_{ST} + 2u}{(8Nu+1)F_{ST} + 4u} \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} N = -\dfrac{F_{IS}+1}{8uF_{IS}} \\ q_s = \dfrac{\left(8\left(-\dfrac{F_{IS}+1}{8uF_{IS}}\right)u+1\right)F_{ST} + 2u}{\left(8\left(-\dfrac{F_{IS}+1}{8uF_{IS}}\right)u+1\right)F_{ST} + 4u} = \dfrac{\left(1-\dfrac{F_{IS}+1}{F_{IS}}\right)F_{ST} + 2u}{\left(1-\dfrac{F_{IS}+1}{F_{IS}}\right)F_{ST} + 4u} \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} N = -\dfrac{F_{IS}+1}{8uF_{IS}} \\ q_s = \dfrac{2u - \dfrac{F_{ST}}{F_{IS}}}{4u - \dfrac{F_{ST}}{F_{IS}}} = \dfrac{2uF_{IS} - F_{ST}}{4uF_{IS} - F_{ST}} \end{cases} \tag{80}$$

Nous savons aussi qu'ici (n = 2 sous-populations) $q_s = 1 - 2m(1 - m)$, soit :

$$q_s = 1 - 2m - 2m^2$$

⇔

$$\frac{q_s}{2} = \frac{1}{2} - m + m^2$$

⇔

$$m^2 - 2\frac{1}{2}m + \left(\frac{1}{2}\right)^2 = \frac{q_s}{2} - \frac{1}{2} + \left(\frac{1}{2}\right)^2$$

⇔

$$m^2 - 2\frac{1}{2}m + \left(\frac{1}{2}\right)^2 = \frac{q_s}{2} - \frac{1}{2} + \left(\frac{1}{2}\right)^2$$

⇔

$$\left(m - \frac{1}{2}\right)^2 = \frac{q_s}{2} - \frac{1}{2} + \left(\frac{1}{2}\right)^2$$

⇔

$$m - \frac{1}{2} = \pm\sqrt{\frac{q_s}{2} - \frac{1}{2} + \left(\frac{1}{2}\right)^2}$$

⇔

$$m = \frac{1}{2} \pm \frac{1}{2}\sqrt{2q_s - 1}$$

Nous avons déjà vu que $m \leq 0,5$ donc :

$$m = \frac{1}{2} - \frac{1}{2}\sqrt{2q_s - 1} \qquad (81)$$

En combinant les équations (80) et (81), nous obtenons :

$$\begin{cases} N = -\dfrac{F_{IS} + 1}{8uF_{IS}} \\ m = \dfrac{1}{2} - \dfrac{1}{2}\sqrt{2\dfrac{2uF_{IS} - F_{ST}}{4uF_{IS} - F_{ST}} - 1} \end{cases}$$

⇔

$$\begin{cases} N = -\dfrac{F_{IS} + 1}{8uF_{IS}} \\ m = \dfrac{1}{2} - \dfrac{1}{2}\sqrt{\dfrac{4uF_{IS} - 2F_{ST} - 4uF_{IS} + F_{ST}}{4uF_{IS} - F_{ST}}} \end{cases}$$

⇔

$$\begin{cases} N = -\frac{F_{IS}+1}{8uF_{IS}} \\ m = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{F_{ST}}{F_{ST}-4uF_{IS}}} \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} N = -\frac{F_{IS}+1}{8uF_{IS}} \\ m = \frac{1}{2}\left[1 - \sqrt{\frac{F_{ST}}{F_{ST}-4uF_{IS}}}\right] \end{cases}$$

Nous pouvons maintenant calculer les effectifs clonaux des différents foyers, ainsi que la proportion de migrants. Notez que dans le cas particulier des clones structurés en deux sous-unités, le F_{IS} devient indépendant de la migration et on peut directement estimer N à partir du F_{IS}. Nous prendrons comme précédemment $u = 0{,}001$. Les résultats sont présentés dans le tableau 38.

Une sous-population isolée

Dans ce cas, on considère que $m = 0$ et donc $q_s = 1$, $q_d = 0$ et $Q_T = 0$ et le système d'équations 73 devient:

$$\begin{cases} Q_I = 0 \\ Q_S = \gamma\left[\frac{1}{2N} + \left(1 - \frac{1}{N}\right)Q_S\right] \\ Q_T = 0 \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} Q_I = 0 \\ Q_S = \gamma\left[\frac{1}{2N} + \left(1 - \frac{1}{N}\right)Q_S\right] \\ Q_T = 0 \end{cases}$$

$\Leftrightarrow$

$$\begin{cases} Q_I = 0 \\ Q_S \frac{2N - 2\gamma(N-1)}{2N} = \frac{\gamma}{2N} \\ Q_T = 0 \end{cases}$$

⇔

$$\begin{cases} Q_I = 0 \\ Q_S \dfrac{2N - 2\gamma(N-1)}{2N} = \dfrac{\gamma}{2N} \\ Q_T = 0 \end{cases}$$

Nous savons que $u \approx 0{,}001$, donc que $u^2 << 1$ et donc que $\gamma \approx 1 - 2u$, ce qui fait :

$$\begin{cases} Q_I = 0 \\ Q_S = \dfrac{1-2u}{2N - 2(1-2u)(N-1)} \\ Q_T = 0 \end{cases}$$

⇔

$$\begin{cases} Q_I = 0 \\ Q_S = \dfrac{1-2u}{2N - 2(1-2u)(N-1)} \\ Q_T = 0 \end{cases}$$

⇔

$$\begin{cases} Q_I = 0 \\ Q_S = \dfrac{1-2u}{2N - 2N + 2 + 4Nu - 4u} = \dfrac{1-2u}{2 + 4Nu - 4u} \\ Q_T = 0 \end{cases}$$

Nous allons considérer que $u << 1$, ce qui fait :

$$\begin{cases} Q_I = 0 \\ Q_S \approx \dfrac{1}{2 + 4Nu} \\ Q_T = 0 \end{cases}$$

Nous ne pouvons ici nous contenter de calculer un F_{IS} :

$$F_{IS} = \frac{Q_I - Q_S}{1 - Q_S} \approx \frac{0 - \frac{1}{2+4Nu}}{1 - \frac{1}{2+4Nu}} \Leftrightarrow F_{IS} \approx \frac{-1}{4Nu+1}$$

$$\Leftrightarrow F_{IS}(4Nu+1) = -1 \Leftrightarrow F_{IS}(4Nu+1) = -1$$

C'est le même résultat que dans l'article de Simo *et al.* (2010). Les résultats de cette approche, en utilisant u = 0,001 comme ailleurs, figurent également sur le tableau 38. Les renseignements complémentaires figurent quant à eux dans le tableau 39 (F_{IS} et F_{ST}).

Tableau 38
Récapitulatif de l'estimation de la taille des populations (*N*) et taux de migration (*m*) de *Trypanosoma brucei gambiense* en Côte d'Ivoire et en Guinée. F_{ST}' provient du calcul décrit auparavant et présenté dans le tableau 37. Les intervalles de confiances à 95 % (Li et Ls) sont obtenus par bootstrap sur les loci sauf pour l'estimation de *m* avec le modèle à deux îles où la méthode du jackknife a été utilisée sur les quatre loci disponibles. Les valeurs de F_{IS} et de F_{ST} utilisées figurent dans le tableau 39.

Méthode	Sous-échantillon	N	Li	Ls	m	Li	Ls	Nm	Li	Ls
F_{ST}'	Boffa-Dubréka							0,21		
Waples	Bonon 2000-2002	32	11	90						
	Bonon 2000-2004	169	69	422						
	Bonon 2002-2004	97	37	287						
	Bonon (moyenne)	100	39	266						
	Dubréka 1998-2002	96	28	342						
MLNE	Bonon	7	6	13	0,365	0,112	0,836	2,71	0,63	11,01
Maximum *likelihood*	Dubréka	5	16	3	0,315	0,918	0,052	1,65	14,66	0,17
MLNE	Bonon	42			0,050			2,12		
Moment	Dubréka	77			0,036			2,72		
Modèle infinité d'îles	Bonon							0,13	0,05	0,22
	Boffa							0,05	0,01	0,10
	Dubréka							0,23	0,10	0,45
Modèle deux îles	Bonon	64	27	109						
	Boffa	25	7	50	0,016	0,005	Infini	0,39	0,26	Infini
	Dubréka	117	49	223	0,010	0,003	Infini	1,15	0,73	Infini
Modèle îles isolées	Bonon	127	53	218						
	Boffa	50	14	100						
	Dubréka	234	98	446						
Moyennes	Bonon	68	31	152	0,207	0,112	0,836	1,65	0,34	5,61
	Boffa	38	10	75	0,016	0,005	Infini	0,22	0,14	0,10
	Dubréka	106	48	254	0,010	0,003	Infini	1,19	5,16	0,31

Il est important de spécifier que pour les méthodes basées sur les différenciations spatiales, temporelles ou spatio-temporelles, c'est un effectif efficace de génotypes multilocus que l'on obtient. Des simulations effectuées avec une version de Easypop modifiée par Franck Prugnolle (disponible sur demande) montrent que dans ce cas, on obtient un N_e très inférieur à N_c (taille de recensement).

Tableau 39
Récapitulatif des valeurs utilisées pour le calcul des effectifs clonaux à partir des modèles. Les F_{IS} ont été calculés avec les données des sept meilleurs loci et en séparant les méthodes d'isolement (plus d'échantillons). Les intervalles de confiance des F_{IS} correspondent aux bootstraps sur les loci, ceux du F_{ST} à un jackknife sur les quatre loci disponibles dans ce cas (trois ne varient pas d'un locus à l'autre et donnent un θ de Weir et Cockerham indéfini.

Sous-échantillon	F_{IS}	**Li**	**Ls**
Bonon	- 0,663	- 0,825	- 0,534
Boffa	- 0,833	- 0,947	- 0,714
Dubréka	- 0,517	- 0,719	- 0,359
	F_{ST}	**Li**	**Ls**
Boffa/Dubréka	0,051	- 0,054	0,156

Dans le cas des estimations effectuées à partir des modèles de populations clonales, c'est un effectif clonal que l'on estime (population clonale d'une taille N_a dérivant à la même vitesse que celle observée), en principe assez proche de N_c sauf si la population n'est pas totalement clonale, auquel cas on risque de surestimer l'effectif réel. Mais ce n'est pas le cas ici comme on l'a vu.

La première chose que l'on remarque est que les effectifs efficaces et clonaux sont du même ordre de grandeur et correspondent assez bien aux nombres de personnes infectées, tels qu'estimés pour chaque foyer dans le tableau 34. Ceci est étonnant, car on sait que les N_e devraient être très petits par rapport au nombre réel de souches présentes. Par ailleurs, comme cela avait été montré dans l'article initial (Koffi *et al.*, 2009), si un taux de mutation de 10^{-4} est utilisé au lieu de 0,001 comme ici, les effectifs clonaux se trouvent multipliés par 10, suggérant une sous-estimation du nombre de souches circulantes estimé par les prospections médicales. Ceci ne devrait cependant pas changer beaucoup l'estimation du *Nm*. Or dans ce cas, nous observons une variation entre 0,2 et 2 individus échangés par génération. Si nous prenons un maximum de 10 générations de trypanosomes par année (comme discuté ailleurs), nous obtenons un maximum d'individus échangés de l'ordre de 2 à 20 par an. Cela signifie, si une éradication séquentielle devait être envisagée (idéalement couplée d'ailleurs à une lutte vectorielle), qu'il faudrait d'abord s'occuper des plus gros foyers (Dubréka en

Guinée), qui envoient le plus de migrants ailleurs, avant de s'occuper des plus petits (Boffa) et ce sans laisser passer trop de temps.

Structure à l'échelle sub-spécifique

Il ne nous reste plus maintenant qu'à étudier comment s'organisent les différentes souches de cette étude entre elles et comment elles se positionnent par rapport à des souches de référence des différentes sous-espèces du complexe *T. brucei*. Nous allons effectuer pour ce faire une analyse NJTree qui est, à mon avis, la plus illustrative. Vous connaissez maintenant la procédure par cœur. En prenant le jeu de données complet, vous le faites passer à la moulinette Create pour obtenir un jeu de données MSA. Avec ce dernier, vous obtenez une matrice de distances harmoniques de Cavalli-Sforza et Edwards entre individus (isolats) que vous faites passer dans MEGA pour dessiner l'arbre. Cet arbre est représenté en figure 98. On peut tout d'abord y voir une certaine disparité avec l'arbre présenté en figure supplémentaire de Koffi *et al.* (2009). Ceci est dû au fait que nous avons utilisé MSA pour calculer les distances harmoniques de Cavalli-Sforza et Edwards. Je me suis aperçu récemment que Genetix ne calcule en fait pas la distance harmonique, mais une version plus ancienne. Ensuite, on peut également remarquer que les souches Tbg1 sont toutes ensemble y compris celles de référence, avec une souche de notre échantillon très divergent par rapport aux autres. Nous remarquons également que les souches de référence Tbg1, qui proviennent du Congo et du Cameroun, se groupent avec les souches de Côte d'Ivoire et jamais avec la Guinée. Ceci est à mettre en parallèle avec la très forte divergence déjà mise en évidence plus haut entre les souches guinéennes et les autres. Enfin, on voit nettement que les autres sous-espèces, Tbb, Tbr ne correspondent à rien de concret génétiquement et qu'en particulier Tbg2, lui-même très hétérogène, n'a aucun rapport génétique avec Tbg1.

CONCLUSION

Après avoir exclu un locus manifestement défectueux, nous avons pu démontrer que la technique d'isolement ne sélectionne pas des génotypes très particuliers en ce qui concerne les génotypes obtenus avec les microsatellites. Il apparaît donc que l'apparente sélection de génotypes isoenzymatiques prend davantage sa source dans la sélection de cellules trypanosomiales à différents stades de développement exprimant différents loci (régulation de l'expression). Cela confirme, si besoin était, que l'utilisation de marqueurs non codants comme les microsatellites est toujours souhaitable pour effectuer des analyses de génétique des populations inférentielle.

L'analyse de l'hétérozygotie relative (F_{IS}) démontre que la recombinaison sexuée est suffisamment rare pour n'avoir laissé aucune signature sur les échantillons examinés.

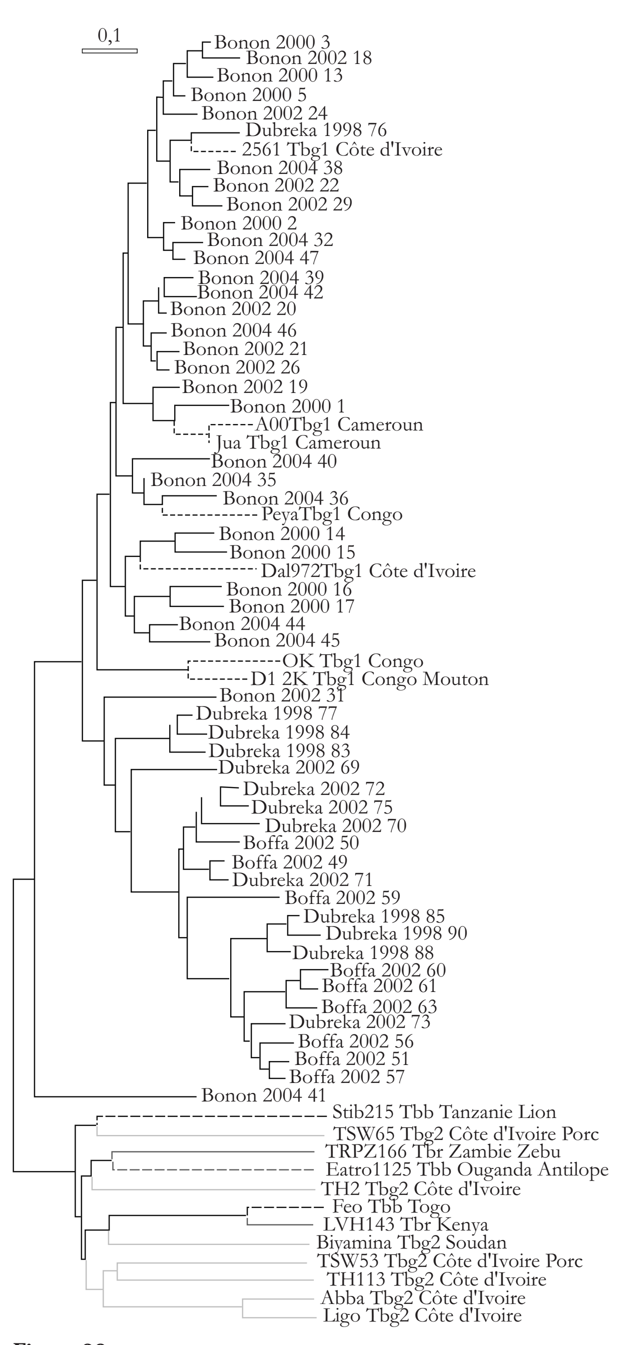


Figure 98
NJTree basé sur la distance harmonique de Cavalli-Sforza et Edwards. Les isolats de l'étude de génétique des populations sont en noir (un seul représentant par MLG pour gagner de la place). Les souches de référence sont en couleur (rouge = Tbg1, jaune = Tbg2, vert = Tbb, bleu = Tbr). L'espèce hôte est indiquée si non-humain.

Les analyses concernant des tailles génétiques des populations de Tbg1 suggèrent un nombre de souches circulantes supérieur à l'estimation du nombre de personnes infectées. Les hôtes réservoirs et/ou les patients asymptomatiques représentent les hypothèses les plus parcimonieuses pour expliquer cela, mais ceci nécessitera confirmation par d'autres types d'études.

La différenciation génétique entre Côte d'Ivoire et Guinée indique une divergence extrême entre ces deux pays. La Guinée semble en effet abriter des souches qui s'éloignent de toutes celles présentes dans notre étude et montrent même des caractéristiques épidémiologiques très différentes des autres (majorité des souches dans les ganglions cérébraux au lieu du sang) (Camara *et al.*, 2005). Nous pouvons ajouter qu'en Guinée le vecteur de la maladie du sommeil est *Glossina palpalis gambiensis*, alors que c'est *G. palpalis palpalis* dans les autres zones concernées par notre étude. Il existe donc vraisemblablement plusieurs taxons distincts au sein de l'entité Tbg1. Que dire alors des autres sous-espèces qui ne se raccrochent a rien ? Que probablement beaucoup reste à faire sur la taxonomie et l'écologie de ces organismes.

Bibliographie

AGAPOW P. M., BURT A., 2001 – Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* 1 : 101-102.

AGRAWAL A., LIVELY C. M., 2002 – Infection genetics: gene-for-gene versus matching alleles models and all points in between. *Evol. Ecol. Res.*, 4 : 79-90.

AKAIKE H., 1974 – A new look at the statistical model identification. *IEEE Trans. Auto. Control*, 19 : 716-723.

AKSOY S., BERRIMAN M., HALL N., HATTORI M., HIDE W., LEHANE M. J., 2005 – A case for a *Glossina* genome project. *Trends Parasitol.*, 21 : 107-111.

ANDERSON E. C., WILLIAMSON E. G., THOMPSON E. A., 2000 – Monte Carlo evaluation of the likelihood for Ne from temporally spaces samples. *Genetics*, 156 : 2109-2118.

ANGERS B., MAGNAN P., PLANTE M., BERNATCHEZ L., 1999 – Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Mol. Ecol.*, 8 : 1043-1053.

ARNAVIEHLE S., DE MEEÛS T., BLANCART A., MALLIÉ M., RENAUD F., BASTIDE J. M., 2000 – Multicentric study of *Candida albicans* isolates from non-neutropenic patients: Population structure and mode of reproduction. *Mycoses*, 43 : 109-117.

ASHLEY Jr C. T., WARREN S. T., 1995 – Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, 29 : 703-728.

AVISE J. C., 2000 – *Phylogegraphy: the History and Formation of Species*. Harvard University Press, Cambridge, Massachusets.

AVISE J. C., ARNOLD J., BALL R. M., BERMINGHAM E., LAMB T., NEIGEL J. E., REEB C. A., SAUNDERS N. C., 1987 – Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Ann. Rev. Ecol. Syst.*, 18 : 489-522.

BADOC C., DE MEEÛS T., BERTOUT S., ODDS F. C., MALLIÉ M., BASTIDE J.-M., 2002 – Clonality structure in *Candida dubliniensis*. *FEMS Microbiol. Let.*, 209 : 249-254.

BALLOUX F., 2001 – EASYPOP (version 1.7): A computer program for population genetics simulations. *J. Hered.*, 92 : 301-302.

BALLOUX F., 2004 – Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution*, 58 : 1891-1900.

BALLOUX F., BRÜNNER H., LUGON-MOULIN N., HAUSSER J., GOUDET J., 2000 – Microsatellites can be misleading: an empirical and simulation study. *Evolution*, 54 : 1414-1422.

BALLOUX F., GOUDET J., 2002 – Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Mol. Ecol.*, 11 : 771-783.

BALLOUX F., LEHMANN L., DE MEEÛS T., 2003 – The population genetics of clonal or partially clonal diploids. *Genetics*, 164 : 1635-1644.

BALLOUX F., LUGON-MOULIN N., 2002 – The estimation of population differentiation with microsatellite markers. *Mol. Ecol.*, 11 : 155-165.

BARNABÉ C., BRISSE S., TIBAYRENC M., 2000 – Population structure and genetic typing of *Trypanosoma cruzi*, the agent of Chagas disease: a multilocus enzyme electrophoresis approach. *Parasitology*, 120 : 513-526.

BARRÉ N., BIANCHI M., CHARDONNET L., 2001 – Role of rusa deer *Cervus timorensis russa* in the cycle

of the cattle tick *Boophilus microplus* in New Caledonia. *Exp. Appl. Acarol.*, 25 : 79-96.

Bartley D., Bagley M., Gall G., Bentley B., 1992 – Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. *Conserv. Biol.*, 6 : 365-375.

Barton D. E., David F. N., 1956 – Some notes on ordered random intervals. *J. Roy. Stat. Soc. Ser. B*, 18 : 79-94.

Belkhir K., Borsa P., Chikhi L., Raufaste N., Bonhomme F., 2004 – GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, université de Montpellier II, Montpellier (France).

Ben Abderrazak S., Guerrini F., Mathieu-Daudé F., Truc P., Neubaeur K., Lewicka K., Barnabé C., Tibayrenc M., 1993 – « Isoenzyme electrophoresis for parasite characterization ». *In* Hyde J. E. (ed.) : *Protocols in Molecular Parasitology*, Humana Press, Totowa, NJ : 361-362.

Bennett J. A., 2004 – « Pest and diseases in the Pacific War: Crossing the line ». *In* Tucker R. P., Russell E. (eds) : *Natural Ennemy, Natural Ally: Toward an Environment History of Warfare*, Oregon State University Press, Corvallis : 217-251.

Benzécri J. P., 1973 – *L´analyse des données.* Tome I. *La taxinomie.* Tome II. *L´analyse des correspondances.* Paris, Dunod.

Borges E. C., Dujardin J. P., Schofield C. J., Romanha A. J., Diotaiuti L., 2000 – Genetic variability of *Triatoma brasiliensis* (Hemiptera: Reduviidae) populations. *J. Med. Entomol.*, 37 : 872-877.

Bougnoux M. E., Aanensen D. M., Morand S., Théraud M., Spratt B. G., d'Enfert C., 2004 – Multilocus sequence typing of *Candida albicans*: strategies, data exchange and applications. *Infect. Genet. Evol.*, 4 : 243-252.

Bouyer J., Balenghien T, Ravel S., Koné N., Vial L., Sidibé I., Solano P., De Meeûs T., 2009 – Population sizes and dispersal pattern of tsetse flies: rolling on the river? *Mol. Ecol.*, 18 : 2787-2797.

Bouyer J., Guerrini L., Desquesnes M., de la Rocque S., Cuisance D., 2006 – Mapping African Animal Trypanosomosis risk from the sky. *Vet. Res.*, 37 : 633-645.

Bowcock A. M., Ruizlinares A., Tomfohrde J., Minch E., Kidd J. R., Cavalli-Sforza L. L., 1994 – High-resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368 : 455-457.

Brenière S. F., Barnabé C., Bosseno M. F., Tibayrenc M., 2003 – Impact of number of isoenzyme loci on the robustness of intraspecific phylogenies using multilocus enzyme electrophoresis: consequences for typing of Trypanosoma cruzi. *Parasitology*, 127 : 273-281.

Brookfield J. F. Y., 1996 – A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol. Ecol.*, 5 : 453-455.

Brown A. H. D., Feldman M. W., Nevo E., 1980 – Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics*, 96 : 523-536.

Brun R., Hecker H., Lun Z. R., 1998 – *Trypanosoma evansi* and *T. equiperdum*: distribution, biology, treatment and phylogenetic relationship (a review). *Vet. Parasitol.*, 79 : 95-107.

Caillaud M. C., Boutin M., Braendle C., Simon J. C., 2002 – A sex-linked locus controls wing polymorphism in males of the pea aphid, *Acyrthosiphon pisum* (Harris). *Heredity*, 89 : 346-352.

Camara M., Kaba D., Kagbadouno M., Sanon J. R., Ouendeno P., Solano P., 2005 – La trypanosomose humaine africaine en zone de mangrove en Guinée : caractéristiques épidémiologiques et cliniques de deux foyers voisins. *Med. Trop.*, 65 : 155-161.

Camara M., Harling Caro-Riaño H., Ravel S., Dujardin J.-P., Hervouet J.-P., de Meeüs T., Kagbadouno M. S., Bouyer J., Solano P., 2006 – Genetic and morphometric evidence for isolation of a tsetse (Diptera: Glossinidae) population (Loos islands, Guinea). *Journal of Medical Entomology*, 43 : 853-860.

Caterino M. S., Cho S., Sperling F. A. H., 2000 – The current state of insect molecular systematics: a thriving tower of Babel. *Annu. Rev. Entommol.*, 45 : 1-54.

Cavalli-Sforza L. L., Edwards A. W. F., 1967 – Phylogenetic analysis: model and estimation procedures. *Am. J. Hum. Genet.*, 19 : 233-257.

Chambers J. M., Hastie T. J., 1992 – Statistical Models in S. Wadsworth and Brooks Cole Advanced Books and Software, PacificGrove, CA.

Chapuis M. P., Estoup A., 2007 – Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.*, 24 : 621-631.

Chessel D., Dufour A. B., Thioulouse J., 2004 –The ade4 package - I: One-table methods. R-News. 4:1.

Chevillon C., Koffi B. B., Barré N., Durand P., Arnathau C.**,** De Meeûs T., 2007a – Direct and indirect inferences on parasite mating and gene transmission patterns. Pangamy in the cattle tick *Rhipicephalus (Boophilus) microplus*. *Infect. Genet. Evol.*, 7 : 298-304.

Chevillon C., Ducornez S., De Meeûs T., Koffi B. B., Gaia H., Delathière J. M., Barré N., 2007b – Accumulation of acaricide resistance mechanisms in *Rhipicephalus (Boophilus) microplus* (Acari: Ixodidae) populations from New Caledonia Island. *Vet. Parasitol.*, 147 : 276-288.

Chippindale A. K., Rice W. R., 2001 – Y chromosome polymorphism is a strong determinant of male fitness in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.*, *USA*, 98 : 5677-5682.

Chlyeh G., Henry P. Y., Sourouille P., Delay B., Khallaayoune K., Jarne P., 2002 – Population genetics and dynamics at short spatial scale in *Bullinus truncatus*, the intermediate host of *Schistosoma haematobium*, in Morocco. *Parasitology*, 125 : 349-357.

Cockerham C. C., 1969 – Variance of gene frequencies. *Evolution*, 23 : 72-84.

Cockerham C. C., 1973 – Analysis of gene frequencies. *Genetics*, 74 : 679-700.

Coombs J. A., Letcher B. H., Nislow K. H., 2008 – CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. *Mol. Ecol. Resour.*, 8 : 578-580.

Corander J., Waldmann P., Sillanpää M. J., 2003 – Bayesian analysis of genetic differentiation between populations. *Genetics*, 163 : 367-374.

Corander J., Waldmann P., Marttinen P., Sillanpää M. J., 2004 – BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20 : 2363-2369.

Corley L. S., Blankenship J. R., Moore A. J., 2001 – Genetic variation and asexual reproduction in the facultatively parthenogenetic cockroach *Nauphoeta cinerea*: implications for the evolution of sex. *J. Evol. Biol.*, 14 : 68-74.

Cornuet J. M., Luikart G., 1996 – Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, 144 : 2001-2014.

Cornuet J. M., Piry S., Luikart G., Estoup A., Solignac M., 1999 – New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, 153 : 1989-2000.

Coustau C., Renaud F., Maillard C., Pasteur N., Delay B., 1991 – Differential susceptibility to a trematode parasite among genotypes of the *Mytilus edulis/galloprovincialis* complex. *Genet. Res. Camb.*, 57 : 207-212.

Cox D. R., Snell E. J., 1981 – *Applied Statistics; Principles and Examples.* London, Chapman and Hall.

Criscione C. D., Blouin M. S., 2005 – Effective sizes of macroparasite populations: a conceptual model. *Trends Parasitol.*, 21 : 212-217.

Criscione C. D., Poulin R., Blouin M. S., 2005 – Molecular ecology of parasites: elucidating ecological and microevolutionary processes. *Mol. Ecol.*, 14 : 2247-2257.

Cutullé C., Jonsson N. N., Seddon J. M., 2010 – Multiple paternity in *Rhipicephalus* (*Boophilus*) *microplus* confirmed by microsatellite analysis. *Exp. Appl. Acarol.*, 50 : 51-58.

David P., Pujol B., Viard F., Castella V., Goudet J., 2007 – Reliable selfing rate estimates from imperfect population genetic data. *Mol. Ecol.*, 16 : 2474-2487.

De Garine-Wichatitsky M., De Meeûs T., Chevillon C., Berthier D., Barre N.,

Thévenon S., Maillard J. C., 2009 – Population genetic structure of wild and farmed rusa deer (*Cervus timorensis russa*) in New-Caledonia inferred from polymorphic microsatellite loci. *Genetica*, 137 : 313-323.

Delaye C., Aeschlimann A., Renaud F., Rosenthal B., De Meeûs T., 1998 – Isolation and characterisation of microsatellite markers in the *Ixodes ricinus* complex (Acari: Ixodidae). *Molec. Ecol.*, 7 : 360-361.

Delaye C., Béati L., Aeschlimann A., Renaud F., De Meeûs T., 1997 – Population genetics structure of *Ixodes ricinus* in Switzerland from allozymic data: No evidence of divergence between nearby sites. *Int. J. Parasitol.*, 27 : 769-773.

De Meeûs T., 2000 – « Adaptive diversity, specialisation, habitat preference and parasites ». *In* Poulin R., Morand S., Skorping A. (eds) : *Evolutionary Biology of Host Parasite Relationships: Theory Meets Reality*, Amsterdam, Elsevier : 27-42.

De Meeûs T., Agnew P., Prugnolle F., 2007b – Asexual Reproduction: Genetics and Evolutionary Aspects. *Cell. Mol. Life Sci.*, 64 : 1355-1372.

De Meeûs T., Balloux F., 2004 – Clonal reproduction and linkage disequilibrium in diploids: a simulation study. *Infect. Genet. Evol.*, 4 : 345-351.

De Meeûs T., Balloux F., 2005 – *F*-statistics of clonal diploids structured in numerous demes. *Mol. Ecol.*, 14 : 2695-2702.

De Meeûs T., Béati L., Delaye C., Aeschlimann A., Renaud F., 2002a – Sex-biased genetic structure in the vector of Lyme disease, *Ixodes ricinus*. *Evolution*, 56 : 1802-1807.

De Meeûs T., Durand P., Renaud F., 2003 – Species concepts: what for? *Trends Parasitol.*, 19 : 425-427.

De Meeûs T., Goudet J., 2000 – Adaptive diversity in heterogeneous environments for populations regulated by a mixture of soft and hard selection. *Evol. Ecol. Res.*, 8 : 981-995.

De Meeûs T., Goudet J., 2007 – A step by step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infect. Genet. Evol.*, 7 : 731-735.

De Meeûs T., Guégan J. F., Teriokhin A., 2009 – MultiTest V.1.2, a program to binomially combine independent tests and performance comparison with other related methods on proportional data. *BMC Bioinformatics*, 10 : 443.

De Meeûs T., Humair P. F., Delaye C., Grunau C., Renaud F., 2004a – Non-Mendelian transmission of alleles at microsatellite loci: an example in *Ixodes ricinus*, the vector of Lyme disease. *Int. J. Parasitol.*, 34 : 943-950.

De Meeûs T., Koffi B. B., Barré N., de Garine-Wichatitsky M., Chevillon C., 2010 – Swift sympatric adaptation of a species of cattle tick to a new deer host in New-Caledonia. *Infect. Genet. Evol.*, 10 : 976-983.

De Meeûs T., Lehmann L., Balloux F., 2006 – Molecular epidemiology of clonal diploids: a quick overview and a short DIY (Do It Yourself) notice. *Infect. Genet. Evol.*, 6 : 163-170.

De Meeûs T., Lorimier Y., Renaud F., 2004b – Lyme borreliosis agents and the genetics and sex of their vector, *Ixodes ricinus*. *Micr. Infect.*, 6 : 299-304.

De Meeûs T., McCoy K. D., Prugnolle F., Chevillon C., Durand P., Hurtrez-Boussès S., Renaud F., 2007a – Population genetics and molecular epidemiology or how to “débusquer la bête”. *Infect. Genet. Evol.*, 7 : 308-332.

De Meeûs T., Michalakis Y., Renaud F., Olivieri I., 1993 – Polymorphism in heterogeneous environments, habitat selection and sympatric speciation: Soft and hard selection models. *Evol. Ecol.*, 7 : 175-198.

De Meeûs T., Renaud F., 2002 – Parasites within the new phylogeny of eukaryotes. *Trends Parasitol.*, 18 : 247-251.

De Meeûs T., Renaud F., Mouveroux E., Reynes J., Galeazzi G., Mallié M., Bastide J. M., 2002b – The genetic structure of *Candida glabrata* populations in AIDS and non-AIDS patients. *J. Clin. Microbiol.*, 40 : 2199-2206.

Dieringer D., Schlötterer C., 2003 – Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes*, 3 : 167-169.

Dobson A. J., 1983 – *An Introduction to Statistical Modelling*. London, Chapman and Hall.

Duchesne P., Turgeon J., 2009 – FLOCK: a method for quick mapping of admixture without source samples. *Molecular Ecology Resources*, 9 : 1333-1344.

Duchesne P., Méthot J., Turgeon J., 2010 – FLOCK 2.0. Département de biologie, université Laval, freely downloadable from http://www.bio.ulaval.ca/no_cache/en/department/professors/professors/professeur/11/13/.

Ducornez S., Barre N., Miller R. J., de Garine-Wichatitsky M., 2005 – Diagnosis of amitraz resistance in *Boophilus microplus* in New Caledonia with the modified Larval Packet Test. *Vet. Parasitol.*, 130 : 285-292.

Ellegren H., 2000 – Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.*, 16 : 551-558.

Ellegren H., 2004 – Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5 : 435-445.

England P. R., Cornuet J. M., Berthier P., Tallmon D. A., Luikart G., 2006 – Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv. Genet.*, 7 : 303-308.

Falk-Vairant J., Guerin P. M., de Bruyne M., Rohrer M., 1994 – Some observation on mating and fertilization in the cattle tick *Boophilus microplus*. *Med. Vet. Entomol.*, 8 : 101-103.

Falush D., Stephens M., Pritchard J. K., 2003 – Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164 : 1567-1587.

FAO, 2000 – Impacts of Trypanosomiasis on African Agriculture. *PAAT technical and Scientific series 2.*

Favre L., Balloux F., Goudet J., Perrin N., 1997 – Female-biased dispersal in the monogamous mammal Crocidura russula: evidence from field data and microsatellite patterns. *Proc. Roy. Soc. London B*, 264 : 127-132.

Fernández-López L., Piñeiro E., Marcos R., Velázquez A., Surrallés J., 2004 – Induction of instability of normal length trinucleotide repeats within human disease genes. *J. Med. Genet.*, 41 : 3-9.

Fisher R. A., 1970 – *Statistical Methods for Research Workers*, 14th Edit. Edinburgh, Oliver and Boyd.

Fontanillas P., Petit E., Perrin N., 2004 – Estimating sex-specific dispersal rates with autosomal markers in hierarchically structured populations. *Evolution*, 58 : 886-894.

Foucaud J., Estoup A., Loiseau A., Rey O., Orivel J., 2010 – Thelytokous parthenogenesis, male clonality and genetic caste determination in the little fire ant: new evidence and insights from the lab. *Heredity*, 105 : 205-212.

Frisch J. E., 1999 – Towards a permanent solution for controlling cattle ticks. *Int. J. Parasitol.*, 29 : 57-71.

Frontier S., 1976 – Étude de la décroissance des valeurs propres dans une analyse en composantes principales : comparaison avec le modèle du bâton brisé. *J. Exp. Mar. Biol. Ecol.*, 25 : 67-75.

Gaffney P. M., 1994 – « Heterosis and heterozygote deficiencies in marine bivalves: more light? » *In* Beaumont A. R. (ed.) : *Genetic and Evolution of Aquatic Organisms*, London, Chapman and Hall : 146-153.

Gallardo J. S., Morales J., 1999 – *Boophilus microplus* (Acari: Ixodidae): preoviposition, oviposition, egg hatching and geotropism. *Bioagro*, 11 : 77-87.

Gandon S., 2002 – Local adaptation and the geometry of host-parasite coevolution. *Ecol. Lett.*, 5 : 246-256.

Gandon S., Capowiez Y., Dubois Y., Michalakis Y., Olivieri I., 1996 – Local adaptation and gene for gene coevolution in a metapopulation model. *Proc. R. Soc. Lond. B*, 263 : 1003-1009.

Gao H., Williamson S., Bustamante C. D., 2007 – A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, 176 : 1635-1651.

Garcia A., Courtin D., Solan P., Koffi M., Jamonneau V., 2006 – Human African trypanosomiasis: connecting parasite and host genetics. *Trends Parasitol.*, 22 : 405-409.

Gerber A. S., Loggins R., Kumar S., Dowling T. E., 2001 – Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *Ann. Rev. Genet.*, 35 : 539-566.

Gibson W., 2007 – Resolution of the species problem in African trypanosomes. *Int. J. Parasitol.*, 37 : 829-838.

Goldstein D. B., Schlötterer C., 1999 – *Microsatellites, Evolution and Applications*. Oxford, Oxford University Press.

Goudet J., 1995 – Fstat version 1.2: a computer program to calculate Fstatistics. *J. Hered.*, 86 : 485-486.

Goudet J., 1999 – An improved procedure for testing the effects of key innovations on rate of speciation. *Am. Nat.*, 153 : 550-555.

Goudet J., 2002 – FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3.2). Available from http://www.unil.ch/izea/softwares/fstat.html. Updated from Goudet (1995).

Goudet J., 2005 – HierFstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes*, 5 : 184-186.

Goudet J., Perrin N., Waser P., 2002 – Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Mol. Ecol.*, 11 : 1103-1114.

Goudet J., Raymond M., De Meeûs T., Rousset F., 1996 – Testing differentiation in diploid populations. *Genetics*, 144 : 1933-1940.

Gubler D. J., 1998 – Resurgent vector-borne diseases as a global health problem. *Emerg. Infect. Dis.*, 4 : 442-450.

Guerrero F. D., Nene V. M., George J. E., Barker S. C., Willadsen P., 2006 – Sequencing a new target genome: the *Boophilus microplus* (Acari: Ixodidae) genome Project. *J. Med. Entomol.*, 43 : 9-16.

Guo S. W., Thompson E. A., 1992 –Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, 48 : 361-372.

Haldane J. B. S., 1954 – An exact test for randomness of mating. *J. Genet.*, 52 : 631-635.

Hardy G. H., 1908 – Mendelian proportions in a mixed population. *Science*, 28 : 49-50.

Hartl D. L., Clark A. G., 1989 – *Principles in Population Genetics, Second Edition*. Sinauer Associates Inc., Sunderland, Massachusetts.

Haubold B., Travisano M., Rainey P. B., Hudson R. R., 1998 – Detecting linkage disequilibrium in bacterial populations. *Genetics*, 150 : 1341-1348.

Hauswaldt J. S., Glenn T. C., 2005 – Population genetics of the diamondback terrapin (*Malaclemys terrapin*). *Mol. Ecol.*, 14 : 723-732.

Healy J. A., 1979 – Analysis of α-Glycerophosphate deshydrogenase variability in the tick *Ixodes ricinus* (Acari: Ixodidae). *Genetica*, 1 : 19-30.

Hedrick P. W., 1999 – Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution*, 53 : 313-318.

Hedrick P. W., 2003 – Hopi Indians, cultural selection, and albinism. *Am. J. Phys. Anthropol.*, 121 : 151-156.

Hedrick P. W., 2005 – A standardized genetic differentiation measure. *Evolution*, 59 : 1633-1638.

Hide M., Bañuls A. L., Tibayrenc M., 2001 – Genetic heterogeneity and phylogenetic status of *Leishmania* (*Leishmania*) *infantum* zymodeme MON-1: epidemiological implications. *Parasitology*, 123 : 425-432.

Hirai H., LoVerde P. T., 1995 – FISH techniques for constructing physical maps on schistosomes chromosomes. *Parasitol. Today*, 11 : 310-314.

Hoffman J. I., Forcada J., Trathan P. N., Amos W., 2007 – Female fur seals show active choice for males that are heterozygous and unrelated. *Nature*, 445 : 912-914.

Hoffman J. I., Matson C. W., Amos W., Loughlin T. R., Bickham J. W., 2006 – Deep genetic subdivision within continuously distributed and highly vagile marine mammal, the Steller's sea lion (*Eumetopias jubatus*). *Mol. Ecol.*, 15 : 2821-2832.

Holm S., 1979 – A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 6 : 65-70.

Hoogstraal H., Aeschlimann A., 1982 – Tick-host specificity. *Mitt Schweiz Entomol Ges*, 55 : 5-32.

Hubbard M. J., Cann K. J., Baker A. S., 1998 – Lyme borreliosis: a tick-born spirochaetal disease. *Rev. Med. Microbiol.*, 9 : 99-107.

Hurtrez-Boussès S., Durand P., Jabbour-Zahab R., Guégan J. F., Meunier C., Bargues M. D., Mas-Coma S., Renaud F., 2004 – Isolation and characterization of microsatellite markers in the liver fluke (*Fasciola hepatica*). *Mol. Ecol. Notes*, 4 : 689-690.

Jamonneau V., Barnabé C., Koffi M., Sané B., Cuny G., Solano P., 2003 – Identification of *Trypanosoma brucei* circulating in a sleeping sickness focus in Côte d'Ivoire: assessment of genotype selection by the isolation method. *Infect. Genet. Evol.*, 3 : 143-149.

Jarne P., Lagoda J. L., 1996 – Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, 11 : 424-429.

Jongejan F., Uilenberg G., 2004 – The global importance of ticks. *Parasitology*, 129 : S3-S14.

Kaba D., Dje N. N., Courtin F., Oke E., Koffi M., Garcia A., Jamonneau V., Solano P., 2006 – The impact of war on the evolution of sleeping sickness in west-central Côte d'Ivoire. *Trop. Med. Int. Health*, 11 : 136-143.

Kalinowski S. T., 2002 – Evolutionary and statistical properties of three genetic distances. *Mol. Ecol.*, 11 : 1263-1273.

Kalinowski S. T., Wagner A. P., Taper M. L., 2006 – ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Mol. Ecol. Notes*, 6 : 576-579.

Kimura M., Ohta T., 1978 – Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. USA,* 75 : 2868-2872.

Kimura M., Weiss G. H., 1964 – The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49 : 561-576.

King J. R., Jackson D. A., 1999 – Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, 10 : 67-77.

Kiszewski A. E., Matuschka F. R., Spielman A., 2001 – Mating strategies and spermiogenesis in ixodid ticks. *Annu. Rev. Entomol.*, 46 : 167-182.

Koffi B. B., De Meeûs T., Barré N., Durand P., Arnathau C., Chevillon C., 2006a – Founder effects, inbreeding and effective sizes in the Southern cattle tick: the effect of transmission dynamics and implications for pest management. *Mol. Ecol.*, 15 : 4603-4611.

Koffi B. B., Risterucci A. M., Joulia D., Durand P., Barré N., De Meeûs T., Chevillon C., 2006b – Characterization of polymorphic microsatellite loci within a young *Boophilus microplus* metapopulation. *Mol. Ecol. Notes*, 6 : 502-504.

Koffi M., De Meeûs T., Bucheton B., Solano P., Camara M., Kaba D., Cuny G., Ayala F. J., Jamonneau V., 2009 – Population genetics of *Trypanosoma brucei gambiense*, the agent of sleeping sickness in Western Africa. *Proc. Natl. Acad. Sci. USA*, 106 : 209-214.

Koffi M., Solano P., Barnabé C., De Meeûs T., Bucheton B., N'Dri L., Cuny G., Jamonneau V., 2007 – Genetic characterisation of *Trypanosoma brucei* ssp. by microsatellite typing: new perspectives for the molecular epidemiology of human African trypanosomosis. *Infect. Genet. Evol.*, 7 : 675-684.

Kumar S., Tamura K, Nei M., 2004 – MEGA3: Integrated software for Molecular Evolutionary

Genetics Analysis and sequence alignment. *Brief. Bioinf.*, 5 : 150-163.

Kunz W., 2002 – When is a parasite species a species? *Trends Parasitol.*, 18 : 121-124.

Labruna M. B., Naranjo V., Mangold A. J., Thompson C., Estrada-Peña A., Guglielmone A. A., Jongejan F., de la Fuente J., 2009 – Allopatric speciation in ticks: genetic and reproductive divergence between geographic strains of *Rhipicephalus* (*Boophilus*) *microplus*. *BMC Evol. Biol.*, 9 : 46.

Latch E. K., Dharmarajan G., Glaubitz J. C., Rhodes O. E., 2006 – Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conserv. Genet.*, 7 : 295-302.

Lawrence M. J., 2000 – Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. *Ann. Bot.*, 85 : 221-226.

Le T. H., Blair D., McManus D. P., 2002 – Mitochondrial genomes of parasitic flatworms. *Trends Parasitol.*, 18 : 206-213.

Leblois R., Estoup A., Rousset F., 2003 – Influence of mutational and sampling factors on the estimation of demographic parameters in a 'continuous' population under isolation by distance. *Mol. Biol. Evol.*, 20 : 491-502.

Leblois R., Rousset F., Estoup A., 2004 – Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, 166 : 1081-1092.

Legendre P., Legendre L., 1998 – *Numerical Ecology*, Second English Edition. Amsterdam, Elsevier, Science B.V.

Lehmann T., Hawley W. A., Kamau L., Fontenille D., Simard F., Collins F. H., 1996 – Genetic differentiation of *Anopheles gambiae* populations from East and West Africa: comparison of microsatellites and allozyme loci. *Heredity*, 77 : 192-208.

Lischer H. E. L., Excoffier L., 2012 – PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28 : 298-299.

Luikart G., Cornuet J. M., 1999 – Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, 151 : 1211-1216.

MacArthur B. H., 1957 – On the relative abundance of bird species. *Proc. Natl. Acad. Sci. USA*, 43 : 293-295.

MacLean L., Odiit M., MacLeod A., Morrison L., Sweeney L., Cooper A., Kennedy P. G. E., Sternberg J. M., 2007 – Spatially and genetically distinct African trypanosome virulence variants defined by host interferon-γ response. *J. Infect. Dis.*, 196 : 1620-1628.

MacLeod A., Tweedie A., Welburn S. C., Maudlin I., Turner C. M. R., Tait A., 2000 – Minisatellite marker analysis of *Trypanosoma brucei*: Reconciliation of clonal, panmictic, and epidemic population genetic structures. *Proc. Natl. Acad. Sci. USA.*, 97 : 13442-13447.

MacLeod A., Tweedie A., McLellan S., Hope M., Taylor S., Cooper A., Sweeney L., Turner C. M. R., Tait A., 2005a – Allelic segregation and independent assortment in *T. brucei* crosses: Proof that the genetic system is Mendelian and involves meiosis (vol. 143, pg 12, 2005). *Mol. Biochem. Parasitol.*, 144 : 131-131.

MacLeod A., Tweedie A., McLellan S., Taylor S., Cooper A., Sweeney L., Turner C. M. R., Tait A., 2005b – Allelic segregation and independent assortment in *T. brucei* crosses: Proof that the genetic system is Mendelian and involves meiosis. *Mol. Biochem. Parasitol.*, 143 : 12-19.

MacLeod A., Tweedie A., McLellan S., Taylor S., Hall N., Berriman M., El-Sayed N. M., Hope M., Turner C. M. R., Tait A., 2005c – The genetic map and comparative analysis with the physical map of *Trypanosoma brucei*. *Nucleic Acids Res.*, 33 : 6688-6693.

MacLeod A., Tweedie A., McLellan S., Taylor S., Hall N., Berriman M., El-Sayed N. M., Hope M., Turner C. M. R., Tait A., 2006 – The genetic map and comparative analysis with the physical map of *Trypanosoma brucei* (vol 33, pg 6688, 2005). *Nucleic Acids Res.*, 34 : 764-764.

Manel S., Gaggiotti O. E., Waples R. S., 2005 – Assignment methods: matching biological questions techniques with appropriate techniques. *Trends Ecol. Evol.*, 20 : 136-142.

Manly B. J. F., 1997 – *Randomization and Monte Carlo methods in biology*, 2nd Edition. London, Chapman & Hall.

Mantel N., 1967 – The detection of disease clustering and a generalized regression approach. *Cancer Res.*, 27 : 209-220.

Maynard-Smith J., Smith N. H., O'Rourke M., Spratt B. G., 1993 – How clonal are bacteria? *Proc. Natl. Acad. Sci. USA*, 90 : 4384-4388.

McCoy K. D., Boulinier T., Tirard C., Michalakis Y., 2003 – Host-dependent genetic structure of parasite populations: differential dispersal of seabird tick host races. *Evolution*, 57 : 288-296.

McCoy K. D., Chapuis E., Tirard C., Boulinier T., Michalakis Y., LeBohec C., LeMaho Y., Gauthier-Clerc M., 2005 – Recurrent evolution of host-specialized races in a globally-distributed ectoparasite. *Proc. Roy. Soc. London B.*, 272 : 2389-2395.

McCullagh P., Nelder J. A., 1989 – *Generalized Linear Models.* London, Chapman and Hall.

Meirmans P. G., 2006 – Using the amova framework to estimate a standardized genetic differentiation measure. *Evolution*, 60 : 2399-2402.

Metropolis N., 1987 – The beginning of the Monte Carlo method. *Los Alamos Science*, 15 : 125-130.

Meunier C., Hurtrez-Boussés S, Jabbour-Zahab R., Durand P., Rondelaud D., Renaud F., 2004a – Field and experimental evidence of preferential selfing in the freshwater mollusc *Lymnaea truncatula* (Gastropoda, Pulmonata). *Heredity*, 92 : 316-322.

Meunier C., Hurtrez-Boussés S, Durand P., Rondelaud D., Renaud F., 2004b – Small effective population sizes in a widespread selfing species, *Lymnaea truncatula* (Gastropoda: Pulmonata). *Mol. Ecol.*, 13 : 2535-2543.

Michalakis Y., Excoffier L., 1996 – A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. *Genetics*, 142 : 1061-1064.

Milgroom M. G., 1996 – Recombination and the multilocus structure of fungal populations. *Ann. Rev. Phytopathol.*, 34 : 457-477.

Morgan A. D., Gandon S., Buckling A., 2005 – The effect of migration on local adaptation in a coevolving host-parasite system. Nature, 437 : 253-256.

Murrel A., Barker S. C., 2003 – Synonymy of *Boophilus* Curtice, 1891 with *Rhipicephalus* Koch, 1844 (Acari : Ixodidae). *Syst. Parasitol.*, 56 : 169-172.

Nadler S. A., 1995 – Microevolution and the genetic structure of parasite populations. *J. Parasitol.*, 81 : 395-403.

Nébavi F., Ayala F. J. , Renaud F., Bertout S., Eholié S., Moussa K., Mallié M., De Meeûs T., 2006 – Clonal population structure and genetic diversity of *Candida albicans* in AIDS patients from Abidjan (Côte d'Ivoire). *Proc. Natl. Acad. Sci. USA*, 103 : 3663-3668.

Nei M., Chesser R. K., 1983 – Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.*, 47 : 253-259.

Niklasson M. Tomiuk J., Parker Jr E. D., 2004 – Maintenance of clonal diversity in *Dipsa bifurcata* (Fallén, 1810) (Diptera: Lonchopteridae). I. Fluctuating seasonal selection moulds long-term coexistence. *Heredity*, 93 : 62-71.

Njiokou F., Nkinin S. W., Grébaut P., Penchenier L., Barnabé C., Tibayrenc M., Herder S., 2004 – An isoenzyme survey of *Trypanosoma brucei* s.l. from the Central African subregion: population structure, taxonomic and epidemiological considerations. *Parasitology*, 128 : 645-653.

Nunney L., Baker A. E. M., 1993 – The Role of Deme Size, Reproductive Patterns, and Dispersal in the Dynamics of T-Lethal Haplotypes. *Evolution*, 47 : 1342-1359.

OHTA T., 1982 – Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA*, 79 : 1940-1944.

OSTERKAMP J., WAHL U., SCHMALFUSS G., HAAS W., 1999 – Host-odour recognition in two tick species is coded in a blend of vertebrate volatiles. *J. Comp. Physiol. A Sens. Neural Behav. Physiol.*, 185 : 59-67.

PAETKAU D., STROBECK C., 1995 – The molecular basis and evolutionary history of a microsatellite null allele in bears. *Mol. Ecol.*, 4 : 519-520.

PAETKAU D., CALVERT W., STIRLING I., STROBECK C., 1995 – Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.*, 4 : 347-354.

PASTEUR N., PASTEUR G., BONHOMME F., CATALAN J., BRITTON-DAVIDIAN J., 1987 – *Manuel technique de génétique par électrophorèse des protéines*. Paris, Lavoisier.

PEEL D., OVENDEN J. R., PEEL S. L., 2004 – NeEstimator Version 1.3: software for estimating effective population size, Queensland Government, Department of Primary Industries and Fisheries, freely downloadable from http://www.dpi.qld.gov.au/fishweb/11629.html.

PEMBERTON J. M., SLATE J., BANCROFT D. R., BARRET J. A., 1995 – Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Mol. Ecol.*, 4 : 249-252.

PIRY S., ALAPETITE A., 2003 – GeneClass 2: A Software for Genetic Assignment and First-Generation Migrant Detection. Freely downlable from http://www1.montpellier.inra.fr/URLB/.

PIRY S., ALAPETITE A., CORNUET J. M., PAETKAU D, BAUDOUIN L., ESTOUP A., 2004 – GeneClass2: a software for genetic assignment and first-generation migrant detection. *J. Hered.,* 95 : 536-539.

PIRY S., LUIKART G., CORNUET J. M., 1999 – BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *J. Hered.*, 90 : 502-503.

POSTIC D., GARNIER M., BARANTON G., 2007 – Multilocus sequence analysis of atypical *Borrelia burgdorferi* sensu lato isolates – Description of *Borrelia californiensis* sp. nov., and genomospecies 1 and 2. *Int. J. Med. Microbiol.*, 297 : 263-271.

PRITCHARD J. K., STEPHENS M., DONNELLY. P., 2000 – Inference of population structure using multilocus genotype data. *Genetics*, 155 : 945-959.

PROUT T., 1981 – A note on the island model with sex-dependent migration. *Theor. Appl. Genet.*, 59 : 327-332.

PRUGNOLLE F., CHOISY M., THÉRON A., DURAND P., DE MEEÛS T., 2004a – Sex-specific correlation between heterozygosity and clone size in the trematode *Schistosoma mansoni*. *Mol. Ecol.*, 13 : 2859-2864.

PRUGNOLLE F., DE MEEÛS T., 2002 – Inferring sex-biased dispersal from population genetic tools: a review. *Heredity*, 88 : 161-165.

PRUGNOLLE F., DE MEEÛS T., 2010 – Apparent high recombination rates in clonal parasitic organisms due to inappropriate sampling design. *Heredity*, 104 : 135-140.

PRUGNOLLE F., DE MEEÛS T., DURAND P., SIRE C., THÉRON A., 2002 – Sex-specific genetic structure in *Schistosoma mansoni*: evolutionary and epidemiological implications. *Mol. Ecol.*, 11 : 1231-1238.

PRUGNOLLE F., DURAND P., THÉRON A., CHEVILLON C., DE MEEÛS T., 2003 – Sex-specific genetic structure: new trends for dioecious parasites. *Trends Parasitol.*, 19 : 171-174.

PRUGNOLLE F., THÉRON A., DURAND P., DE MEEÛS T., 2004b – Test of pangamy by genetic analysis of *Schistosoma mansoni* pairs within its natural murine host in Guadeloupe. *J. Parasitol.*, 90 : 507-509.

PRUGNOLLE F., THÉRON A., POINTIER J. P., JABBOUR-ZAHAD R., JARNE P., DURAND P., DE MEEÛS T., 2005 – Dispersal in a parasitic worm and its two hosts and its consequences for local adaptation. *Evolution*, 59 : 296-303.

QIU W. G., BOSLER E. M., CAMPBELL J. R., UGINE G. D., WANG I. N., LUFT B. J., DYKHUIZEN

D. E., 1997 – A population genetic study of *Borrelia burgdorferi sensu stricto* from eastern Long Island, New York, suggested frequency-dependent selection, gene flow and host adaptation. *Hereditas*, 127 (1997) : 203-216.

Queller D. C., Goodnight K. F., 1989 – Estimating relatedness using genetic markers. *Evolution*, 43 : 258-275.

Rageau J., Vergent G., 1959 – Les tiques (Acariens : Ixodidae) des îles françaises du Pacifique. *Bull. Soc. Pathol. Exot.*, 52 : 819-835.

Rannala B., Mountain J. L., 1997 – Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci. USA*, 94 : 9197-9221.

Raufaste N., Bonhomme F., 2000 – Properties of bias of two multiallelic estimators of F_{ST}. *Theor. Pop. Biol.*, 57 : 285-296.

Ravel S., De Meeus T., Dujardin J. P., Zézé D. G., Gooding R. H., Dusfour I., Sané B., Cuny G., Solano P., 2007 – The tsetse fly *Glossina palpalis palpalis* is composed of several genetically differentiated small populations in the sleeping sickness focus of Bonon, Côte d'Ivoire. *Infect. Genet. Evol.*, 16 : 116-125.

Raymond M., Rousset F., 1995a – An exact test for population differentiation. *Evolution*, 49 : 1280-1283.

Raymond M., Rousset F., 1995b – GENEPOP (version .2): population genetics software for exact tests and ecumenicism. *J. Hered.*, 86 : 248-249.

Raymond M., Rousset F., 2003 – GENEPOP (version 3.4): population genetics software for exact tests and ecumenicism (updated from Raymond et Rousset, 1995b).

Razakandrainibe F. G., Durand P., Koella J. C., De Meeüs T., Rousset F., Ayala F. J., Renaud F., 2005 – "Clonal" population structure of the malaria agent *Plasmodium falciparum* in high-infection regions. *Proc. Natl. Acad. Sci. USA*, 102 : 17388-17393.

R-Development-core-team, 2011 – R: A Language and Environment for Statistical Computing. *In* : R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0.

Rice W. R., 1989 – Analyzing tables of statistical tests. *Evolution*, 43 : 223-225.

Richter D., Postic D., Sertour N., Livey I., Matuschka F. R., Baranton G., 2006 – Delineation of *Borrelia burgdorferi sensu lato* species by multilocus sequence analysis and confirmation of the delineation of *Borrelia spielmanii* sp. nov. *Int. J. Syst. Evol. Microbiol.*, 56 : 873-881.

Ridley M., 1996 – *Evolution, Second Edition*. Cambridge, Massachusetts, Blackwell Science, Inc.

Robertson A., Hill W. G., 1984 – Deviations from Hardy-Weinberg proportions: sampling variances and usent in estimation of inbreeding coefficients. *Genetics*, 107 : 713-718.

Roderick G. K., 1996 – Geographic structure of insect populations: gene flow, phylogeography, and their uses. *Annu. Rev. Entomol.*, 41 : 325-352.

Roditi I., Furger A., Ruepp S., Schurch N., Butikofer P., 1998 – Unravelling the procyclin coat of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.*, 91 : 117-130.

Rostand E., 1908 – *Chantecler*. Paris, réédité en 2000 par L'Harmattan.

Rougeron V., De Meeûs T., Hide M., Waleckx E., Bermudez H., Arevalo A., Llanos-Cuentas A., Dujardin J. C., De Doncker S., Le Ray D., Ayala F. J., Bañuls A. L., 2009 – Extreme inbreeding in *Leishmania braziliensis*. *Proc. Natl. Acad. Sci. USA*, 106 : 10224-10229.

Rousset F., 1996 – Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics*, 142 : 1357-1362.

Rousset F., 1997 – Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, 145 : 1219-1228.

Rousset F., 2000 – Genetic differentiation between individuals. *J. Evol. Biol.*, 13 : 58-62.

Rousset F., 2004 – *Genetic Structure and Selection in Subdivided Populations*. Princeton, Princeton University Press.

Rousset F., 2008 – GENEPOP ‘ 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, 8 : 103-106.

Rousset F., Raymond M., 1995 – Testing heterozygote excess and deficiency. *Genetics*, 140 : 1413-1419.

Rousset F., Raymond M., 1997 – Statistical analyses of population genetic data: New tools, old concepts. *Trends Ecol. Evol.*, 12 : 313-317.

Schierup M. H., Mikkelsen A. M., Hein J., 2001 – Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. *Genetics*, 159 : 1833-1844.

Schofield C. J., Kabayo J. P., 2008 – Trypanosomiasis vector control in Africa and Latin America. *Parasit. Vect.*, 1 : 24.

Shaw C. R., 1970 – How many genes evolve? *Bioch. Genet.*, 4 : 275-283.

She J. X., Autem M., Kotoulas G., Pasteur N., Bonhomme F., 1987 – Multivariate analysis of genetic exchanges between *Solea aegyptiaca* and *Solea senegalensis* (Teleosts, Soleidae). *Biol. J. Linn. Soc.*, 32 : 357-371.

Shuker D. M., Reece S. E., Whitehorn P. R., West S. A., 2004 – Sib-mating does not lead to facultative sex ratio adjustment in the parasitoid wasp, *Nasonia vitripennis*. *Evol. Ecol. Res.*, 6 : 73-480.

Siegel S., Castellan Jr. N. J., 1988 – *Nonparametric Statistics for the Behavioral Sciences, Second Edition*. New-York, McGraw-Hill Inc.

Simo G., Njiokou F., Tume C., Lueong S., De Meeûs T., Cuny G., Asonganyi T., 2010 – Population genetic structure of Central African *Trypanosoma brucei gambiense* isolates using microsatellite DNA markers. *Infect. Genet. Evol.*, 10 : 68-76.

Škalamera J. P., Renaud F., Raymond M., De Meeûs T., 1999 – No evidence for genetic differentiation of the mussel *Mytilus galloprovincialis* between lagoons and the seaside. *Mar. Ecol. Prog. Ser.*,178 : 251-258.

Slatkin M., 1985 – Gene flow in natural populations. *Ann. Rev. Ecol. Syst.*, 16 : 393-430.

Slatkin M., 1995 – A measure of population subdivision based on microsatellite allele frequency. *Genetics*, 139 : 457-462.

Šnabel V., De Meeûs T., Varady M., Nansen P.; Bjorn H., Corba J., 2000 – The sexually linked Mpi locus is presumably involved in imidothiazole resistance in *Oesophagostomum dentatum* parasites. *Parasitol. Res.*, 86 : 486-490.

Sokal R. R., Rohlf F. J., 1981 – *Biometry, 2nd Ed.* New-York, Freeman and Co.

Solano P., de La Rocque S., De Meeûs T., Cuny G., Duvallet G., Cuisance D., 2000 – Microsatellite DNA markers reveal genetic differentiation among populations of *Glossina palpalis gambiensis* collected in the agropastoral zone of Sideradougou, Burkina Faso. *Insect. Mol. Biol.*, 9 : 433-439.

Sunnucks P., 2000 – Efficient genetic markers for population biology. *Trends Ecol. Evol.*, 15 : 199-203.

Tabachnick W. J., Black W. C., 1995– Making a Case for Molecular Population Genetic-Studies of Arthropod Vectors. *Parasitol. Today*, 11 : 27-30.

Tait A., MacLeod A., Tweedie A., Masiga D., Turner C. M. R., 2007 – Genetic exchange in *Trypanosoma brucei*: Evidence for mating prior to metacyclic stage development. *Mol. Biochem. Parasitol.*, 151 : 133-136.

Takezaki N, Nei M., 1996 – Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, 144 : 389-99.

Tamura K., Peterson N., Stecher G., Nei M., Kumar S., 2011a – MEGA version 5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods, freely downloadable from http://www.megasoftware.net/.

Tamura K., Peterson N., Stecher G., Nei M., Kumar S., 2011b – MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum

Parsimony Methods. *Mol. Biol. Evol.*, 28 : 2731-2739.

Taylor J. W., Geiser D. M., Burt A., Koufopanou V., 1999 – The evolutionary biology and population genetics underlying fungal strain typing. *Clin. Microbiol. Rev.*, 12 : 126-146.

Terblanche J. S., Chown S. L., 2007 – Factory flies are not equal to wild flies. *Science*, 317 : 1678.

Ter Braak C. J. F., 1986 – Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 : 1167-179.

Ter Braak C. J. F., 1987 – *CANOCO - a Fortran program for canonical community ordination*. Microcomputer Power, Ithaca, New York, USA.

Ter Braak C. J. F., Šmilauer P. 2002 – *CANOCO Reference Manual and CanoDraw for Widows User's Guide: Software for Canonical Community Ordination (version 4.5*). Microcomputer Power, Ithaca, New-York.

Teriokhin A. T., De Meeûs T., Guégan J. F., 2007 – On the power of some binomial modifications of the Bonferroni multiple test. *Zh. Obshch. Biol. (J. Gener. Biol.)*, 68 : 332-340.

Thomas F., Renaud F., Derothe J. M., Lambert A., De Meeûs T., Cezilly F., 1995 – Assortative pairing in *Gammarus insensibilis* (Amphipoda) infested by a trematode parasite. *Oecologia*, 104 : 259-264.

Tibayrenc M., 1998 – Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. *Int. J. Parasitol.*, 28 : 85-104.

Tibayrenc M., 1999 – Toward an integrated genetic epidemiology of parasitic protozoa and other pathogens. *Ann. Rev. Genet.*, 33 : 449-477.

Tibayrenc M., Ayala F. J., 2002 – The clonal theory of parasitic protozoa: 12 years on. *Trends Parasitol.*, 18 : 405-410.

Tibayrenc M., Kjellberg F., Ayala F. J., 1990 – A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci. USA*, 87 : 2414-2418.

Tibayrenc M., Kjellberg F., Arnaud J., Oury B., Brenière S. F., Dardé M. L., Ayala F. J., 1991 – Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proc. Natl. Acad. Sci. USA*, 88 : 5129-5133.

Toonen R. J., 1997 – *Microsatellites for Ecologists: Non-Radioactive Isolation and Amplification Protocols for microsatellite markers*. Unpublished manuscript, available from the author or *via* anonymous FTP from http:/biogeek.ucdavis.edu/Msats/ or http://www2.hawaii.edu/~toonen/files/MsatsV1.pdf.

Trouvé S., Degen L., Goudet J., 2005 – Ecological components and evolution of selfing in the freshwater snail *Galba truncatula*. *J. Evol. Biol.*, 18 : 358-370.

Ueti M. W., Palmer G. H., Scoles G. A., Kappmeyer L. S., Knowles D. P., 2008 – Persistently infected horses are reservoirs for intrastadial tick-borne transmission of the apicomplexan parasite *Babesia equi*. *Infect. Immun.*, 76 : 3525-3529.

Uilenberg G., 1976 – Tick-borne livestock diseases and their vectors. 2. Epizootiology of tick-borne diseases. *World Animal Review*, 17 : 8-15.

Van Bekkum M., Sagar P. M., Stahl J. C., Chambers G. K., 2006 – Natal philopatry does not lead to population genetic differentiation in Buller's albatross (*Thalassarche bulleri bulleri*). *Mol. Ecol.*, 15 : 73-79.

Van Oosterhout C., Hutchinson W. F., Wills D. P. M., Shipley P., 2004 – Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes*, 4 : 535-538.

Verges J., 1944 – *Les tiques du bétail. Méthodes d'éradication*. Nouméa, Imprimeries réunies.

Vignal A., Milan D., SanCristobal M., Eggen A., 2002 – A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.*, 34 : 275-305.

Vitalis R., Couvet D, 2001a – ESTIM 1.0: a computer program to infer population parameters from one- and two-locus gene identity probabilities. *Mol. Ecol. Notes*, 1 : 354-356.

Vitalis R., Couvet D, 2001b – Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, 157 : 911-925.

Vitalis R., Couvet D, 2001c – Two-locus identity probabilities and identity disequilibrium in a partially selfing population. *Genet. Res.*, 77 : 7-81.

Wahlund S., 1928 – Zusammensetzung von populationen und korrelationsers-chinungen von standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11 : 65-108.

Wang J., 2002 – An estimator for pairwise relatedness using molecular markers. *Genetics*, 160 : 1203-1215.

Wang J., Whitlock M. C., 2003 – Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163 : 429-446.

Waples R. S., 1989 – A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121 : 379-391.

Waples R. S., 2006 – A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv. Genet.*, 7 : 167-184.

Waples R. S., Do C., 2008 – LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Mol. Ecol. Res.*, 8 : 753-756.

Waser P., Strobeck C., 1998 – Genetic signatures of interpopulation dispersal. *Trends Ecol. Evol.*, 13 : 43-44.

Wattier R., Engel C. R., Saumitou-Laprade P., Valero M., 1998 – Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). *Mol. Ecol.*, 7 : 1569-1573.

Watts P. C., Rousset F., Saccheri I. J., Leblois R., Kemp S. J., Thompson D. J., 2007 – Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of “neighbourhood size” using a more precise estimator. *Mol. Ecol.*, 16 : 737-751.

Wedekind C., Penn D., 2000 – MHC genes, body odours, and odour preferences. *Nephrol. Dial. Transplant.*, 15 : 1269-1271.

Weinberg W., 1908 – Über den Nachweis der Verebung beim Menschen. Jahresh. *Verein f. Vaterl. Naturk in Wüttemberg*, 64 : 368-382.

Weir B. S., 1979 – Inferences about linkage disequilibrium. *Biometrics*, 35 : 235-254.

Weir B. S., 1996 – *Genetic Data Analysis*. Sinauer Associates Inc., Sunderland, Massachusetts.

Weir B. S., Cockerham C. C., 1984 – Estimating F-statistics for the analysis of population structure. *Evolution*, 38 : 1358-1370.

Whitlock M. C., 2005 – Combining probability from independent tests: the weighted *Z*-method is superior to Fisher’s approach. *J. Evol. Biol.* 18 : 1368-1373.

Whitlock M. C., McCauley D. E., 1998 – Indirect measures of gene flow and migration: $FST \neq 1/(4Nm + 1)$. *Heredity*, 82 : 117-125.

WHO, 2006a – Human African trypanosomiasis (sleeping sickness): epidemiological update. *Weekly Epidemiological Record*, 82 : 71-80.

WHO, 2006b – African trypanosomiasis (sleeping sickness), WHO Media centre, Fact sheet N° 259, World Health Organization, http://www.who.int/mediacentre/factsheets/fs259/en/.

Wilson A. J., Hutchings J. A., Ferguson M. M., 2004 – Dispersal in a stream dwelling salmonid: inferences from tagging and microsatellite studies. *Conserv. Genet.*, 5 : 25-37.

Wolff K. E., 1996 – « Comparison of graphical data analysis methods ». *In* Faulbaum F., Bandilla W. (eds) : *SoftStat ’95 Advances in Statistical Software 5*, Lucius & Lucius, Stuttgart : 139-151.

Wright S., 1951 – The genetical structure of populations. *Ann. Eugenics*, 15 : 323-354.

Wright S., 1965 – The interpretation of population structure by F-statistics with special regard to system of mating. *Evolution*, 19 : 395-420.

Xu J., 2005 – The inheritance of organelle genes and genomes: patterns and mechanisms. *Genome*, 48 : 951-958.

Réponses aux questions

Réponse 1 : L'hypothèse faite est que l'échantillonnage ne modifie pas les fréquences alléliques dans la population, ce qui suppose que cette dernière est suffisamment grande.

Réponse 2 : Les taux de mutation d'une base vers une autre ne sont pas identiques selon que l'on s'adresse à une transversion ou à une transition (voir le glossaire). Une telle propriété interférera nécessairement avec les effets d'ordre démographique. D'ailleurs, la différence est telle qu'on considère en général qu'un site variable ou SNP ne possède que deux allèles possibles A/G ou C/T.

Réponse 3 : Si $s = 1$ alors nous obtenons pour les homozygotes 1/1, les hétérozygotes 1/2 et les homozygotes 2/2, en se souvenant que $p_1 + p_2 = 1$, des fréquences génotypiques respectivement égales à :

$$D_e = p_1^2 + p_1 p_2 \frac{1}{2-1} = p_1^2 + p_1 p_2 = p_1(p_1 + p_2) = p_1$$

$$H_e = 2p_1 p_2\left(1 - \frac{1}{2-1}\right) = 2p_1 p_2 (1-1) = 0$$

$$R_e = p_2^2 + p_1 p_2 \frac{1}{2-1} = p_2^2 + p_1 p_2 = p_2(p_1 + p_2) = p_2$$

Réponse 4 :

$$\overline{(p_i - \overline{p})^2} = \frac{1}{n}\sum_i (p_i - \overline{p})^2 = \frac{1}{n}\sum_i (p_i^2 + \overline{p}^2 - 2p_i\overline{p})$$

$$= \frac{1}{n}\sum_i p_i^2 + \frac{1}{n}\sum_i \overline{p}^2 - \frac{2\overline{p}}{n}\sum_i p_i$$

Et donc

$$\overline{(p_i - \overline{p})^2} = \overline{p^2} + \frac{n}{n}\overline{p}^2 - 2\overline{p}^2 = \overline{p^2} - \overline{p}^2 \text{ (CQFD).}$$

Réponse 5 : C'est la définition de la variance.

Réponse 6 : Dans un modèle en îles infini, s'il n'existe que des sous-populations fixées pour l'un des deux allèles présents à un locus, cela signifie que pour une proportion $\overline{p}$ de populations nous avons $p = 1$ et pour $1 - \overline{p}$ nous avons $p = 0$. La variance de p dans ce cas sera égale à :

$$\sigma_{\max}^2(p) = \frac{1}{n}\sum_i (p_i - \overline{p})^2 = \frac{1}{n}\sum_i \left[n\overline{p}(1-\overline{p})^2 + n(1-\overline{p})(0-\overline{p})^2\right] = \overline{p}(1-\overline{p})^2 + \overline{p}^2(1-\overline{p})$$

d'où l'on tire facilement :

$$\sigma^2_{\max}(p) = \overline{p}(1-\overline{p})(1-\overline{p}+\overline{p}) = \overline{p}(1-\overline{p}) \text{ (CQFD)}$$

Réponse 7 : Détails du calcul d'un *G*

Supposons que nous avons échantillonné N individus dans deux localités différentes (échantillons 1 et 2 de tailles respectives N_1 et N_2). Ces individus ont été génotypés pour un locus qui présente deux allèles de fréquences p_1 et q_1 dans l'échantillon 1 et p_2 et q_2 dans l'échantillon 2 respectivement. Ces informations nous donnent les effectifs d'allèles suivants :

	Nombres observés d'allèles		
Échantillons	**Allèle 1**	**Allèle 2**	**Somme**
Échantillon 1	$2N_1p_1$	$2N_1q_1$	$2N_1(p_1 + q_1) = 2N_1$
Échantillon 2	$2N_2p_2$	$2N_2q_2$	$2N_2(p_2 + q_2) = 2N_2$
Somme	$2N_1p_1 + 2N_2p_2$	$2N_1q_1 + 2N_2q_2$	$2(N_1 + N_2) = 2N$

Si on considère que les individus des deux échantillons proviennent d'une seule et même population (pas de différence réelle de leurs fréquences alléliques) alors, la meilleure estimation de la fréquence des allèles dans la population correspond à la moyenne des fréquences des deux échantillons. Par conséquent, les effectifs attendus des allèles deviennent :

	Effectifs attendus des allèles		
Échantillons	**Allèle 1**	**Allèle 2**	**Somme**
Échantillon 1	$\frac{2N_1p_1 + 2N_2p_2}{2N}2N_1$	$\frac{2N_1p_1 + 2N_2p_2}{2N}2N_1$	$2N_1$
Échantillon 2	$\frac{2N_1p_1 + 2N_2p_2}{2N}2N_2$	$\frac{2N_1p_1 + 2N_2p_2}{2N}2N_2$	$2N_2$
Somme	$2N_1p_1 + 2N_2p_2$	$2N_1q_1 + 2N_2q_2$	$2(N_1 + N_2) = 2N$

Soit P_{MO} la probabilité multinomiale d'observer les effectifs du premier tableau si les fréquences alléliques de chaque échantillon sont correctes et P_{ME} la probabilité multinomiale d'observer ces effectifs si ce sont les effectifs attendus qui sont corrects :

$$P_{MO}=\frac{2N!}{2N_1p_1!2N_1q_1!2N_2p_2!2N_2q_2!}\left(\frac{2N_1p_1}{2N}\right)^{2N_1p_1}\left(\frac{2N_1q_1}{2N}\right)^{2N_1q_1}\left(\frac{2N_2p_2}{2N}\right)^{2N_2p_2}\left(\frac{2N_2q_2}{2N}\right)^{2N_2q_2}$$

$$P_{ME}=\frac{2N!}{2N_1p_1!2N_1q_1!2N_2p_2!2N_2q_2!}$$

$$\left[\frac{2N_1(2N_1p_1+2N_2p_2)}{(2N)^2}\right]^{2N_1p_1}\left[\frac{2N_1(2N_1q_1+2N_2q_2)}{(2N)^2}\right]^{2N_1q_1}$$

$$\times\left[\frac{(2N_1p_1+2N_2p_2)N_2}{(2N)^2}\right]^{2N_2p_2}\left[\frac{(2N_1q_1+2N_2q_2)N_2}{(2N)^2}\right]^{2N_2q_2}$$

Le ratio du logarithme népérien de la vraisemblance ou G correspond à deux fois le logarithme népérien du ratio de vraisemblance, soit :

$G = 2\ln(P_{MO}/P_{ME})$, ce qui peut s'écrire (cf page 736 et Box 17.6 dans SOKAL et ROHLF, 1981) :

$G = 2N_1p_1 \ln(2N_1p_1) + 2N_1q_1 \ln(2N_1q_1) + 2N_2p_2 \ln(2N_2p_2) + 2N_2q_2 \ln(2N_2q_2) + 2N \ln(2N) - 2N_1 \ln(2N_1) - (2N_1p_1 + 2N_2p_2) \ln(2N_1p_1 + 2N_2p_2) - (2N_1q_1 + 2N_2q_2) \ln(2N_1q_1 + 2N_2q_2) - 2N_2 \ln(N_2)$.

Cette quantité possède des propriétés additives, ce qui signifie que les différents G calculés pour différents loci peuvent s'additionner, permettant ainsi d'obtenir un G global offrant donc la possibilité d'un test global.

Réponse 8 : Détails du test de Mantel

Soit M_1 et M_2 deux matrices de distances entre les mêmes paires d'objets :

$$M_1=\begin{bmatrix} m1_{11} & m1_{12} & m1_{13} & m1_{14} \\ & m1_{22} & m1_{23} & m1_{24} \\ & & m1_{33} & m1_{34} \\ & & & m1_{44} \end{bmatrix} \text{ et } M_2=\begin{bmatrix} m2_{11} & m2_{12} & m2_{13} & m2_{14} \\ & m2_{22} & m2_{23} & m2_{24} \\ & & m2_{33} & m2_{34} \\ & & & m2_{44} \end{bmatrix}$$

Une mesure de la corrélation entre ces deux matrices peut par exemple être fournie par :

$$Z=\sum_i\sum_j m1_{ij}\,m2_{ij}$$

Z peut alors être utilisé comme statistique du test de Mantel. Il s'agit de randomiser un grand nombre de fois (10^6 pour Genepop) les objets contenus dans une des deux matrices en mesurant le Z entre la matrice randomisée et l'autre matrice (non randomisée), pour chaque randomisation. La valeur observée du Z peut ensuite être comparée à la distribution des Z randomisés. D'autres statistiques, telles que le classique coefficient de corrélation de Pearson ou, comme dans Genepop, le coefficient de corrélation de rang de Spearman, peuvent également être utilisées à la place du Z pour le test de Mantel.

Réponse 9 : Le critère du bâton brisé ou « broken stick ».

Ce critère a été développé en premier lieu par des écologistes soucieux de comparer la répartition des espèces avec une répartition aléatoire (Barton et David, 1956 ; MacArthur, 1957 pour les premiers). Il fut ensuite adapté aux analyses en composantes principales par Frontier (1976). Selon ce principe, une quantité donnée 1 (correspondant à 100 % de l'information) peut être assimilée à un bâton que l'on peut subdiviser en S parties en y pratiquant S - 1 coupures au hasard. Si ces coupures se font au hasard, on peut avoir n'importe quelle longueur de ces différentes parties avec une probabilité d'apparition qui doit suivre une loi uniforme. Si on classe ces bouts de bois de la plus grande longueur à la plus petite, sous l'hypothèse nulle la plus petite longueur possible sera de $1/S$ avec une probabilité d'apparition de $1/S$. La seconde plus petite sera de longueur $1/S + 1/(S - 1)$ et pour une longueur quelconque l_i on aura :

$$E(l_j) = \frac{1}{S}\sum_{i=0}^{S-j}\frac{1}{j+1}$$

On obtient ainsi la liste par ordre décroissant des espérances de la proportion de variance expliquée par chaque axe sous l'hypothèse nulle. Par exemple, s'il y a 15 axes cela donne la suite 0,221, 0,155, 0,121, 0,099, 0,082, 0,069, 0,058, 0,048, 0,040, 0,033, 0,026, 0,020, 0,014, 0,009, 0,004 qui donne donc les proportions minimales à partir desquelles les axes sont significatifs. Ici, une ACP avec 15 allèles (donc 15 axes) dont le premier axe aurait une inertie inférieure à 22,1 % n'aurait donc aucun axe significatif selon le critère du bâton brisé. Par contre, si cette ACP donne les deux premiers axes avec des inerties (par exemple) de 25 % et 22 % d'inerties suivies d'axes à l'inertie inférieure à 12 %, on a deux axes significatifs selon le critère du bâton brisé.

Réponse 10 : Effectif efficace d'une population dioïque

Soit N_f et N_m, le nombre de mâles et de femelles dans une grande population par ailleurs isolée, sans mutation ni sélection, à générations non chevauchantes, avec accouplements aléatoires (pangamie) et constance du sexe-ratio d'une génération à l'autre. Dans une telle population, pour que deux gènes d'un zygote de la génération t soit formé par deux gènes issus d'un même gène ancêtre, il est nécessaire que ce gène soit présent chez la mère et le père de ce zygote, c'est-à-dire s'il a été prélevé deux fois chez le même mâle de la génération t-2 ou la même femelle de la génération t-2. Sachant qu'il y a pangamie, la probabilité que les deux gènes d'un zygote proviennent du même grand-père est de $1/N_m$ et de la même grand-mère de $1/N_f$. Dans chacun des deux cas, la probabilité de tirer deux fois le même gène chez le grand-parent pour le transmettre aux deux parents du zygote est de $(1/2)^2$ et la probabilité de retirer deux fois ce gène chez le père et la mère pour le transmettre au zygote est aussi de $(1/2)^2$, donc $(1/N_f)\times(1/2)^2\times(1/2)^2$ pour le gène de la grand-mère et $1/16N_m$ pour le gène du grand-père. Les individus étant diploïdes, cet événement

possède deux chances de se réaliser (ou deux essais possibles). La probabilité pour un individu donné que deux gènes pris au hasard découlent d'un même gène ancêtre (coalescence) est donc égale à :

$$\tau = 2\left(\frac{1}{16N_m} + \frac{1}{16N_f}\right) = \frac{N_f + N_m}{8N_f N_m}$$

Nous recherchons l'effectif efficace N_e tel qu'une population monoïque de cette taille dérive à la même vitesse (même coalescence) que notre population dioïque. Sachant que pour une population monoïque, la probabilité de tirer deux fois le même gène est égale à $\tau_e = 1/(2N_e)$, on cherche donc N_e tel que $\tau_e = \tau$, soit :

$$\tau = \frac{1}{2N_e} = \frac{N_f + N_m}{8N_f N_m}$$

⇔

$$N_e = \frac{4N_f N_m}{N_f + N_m} \text{ (CQFD)}$$

Réponse 11 : Estimer un taux de croisements frère-sœur à partir du F_{IS}

Cette méthode a déjà été utilisée dans Chevillon *et al.* (2007a). Si on observe l'évolution de la consanguinité F entre la génération t-2 et t dans une population où les croisements ne se font qu'entre frères et sœurs, on obtient l'image suivante (fig. 99). On cherche à exprimer la consanguinité d'un individu de la génération t, c'est-à-dire que l'on recherche avec quelle probabilité cet individu aura deux gènes identiques par ascendance (issus d'un seul gène ancêtre). Les deux gènes d'un individu pourront être identiques parce qu'ils proviennent du même grand-parent et que ce dernier aura donné deux fois le même gène ou un gène différent, mais déjà identique par ascendance. Ils pourront aussi être identiques s'ils proviennent des deux grands-parents si ces derniers ont des gènes identiques par ascendance. Comme on peut le voir dans la figure 99, la constitution génétique d'un individu de la génération t peut suivre 16 événements différents et équiprobables. Dans la moitié des cas, les deux gènes d'un tel individu proviennent d'un même grand-parent et dans l'autre moitié des cas d'un des deux grands-parents. Quand les deux gènes proviennent d'un même grand-parent, la probabilité de prélever deux fois ce même gène est de $(1/2)^2$ pour le premier gène et la même chose pour le second, soit $P_{2 \text{ mêmes/même grand-parent}} = 1/2$, et celle de prélever les deux gènes différents est aussi $P_{2 \text{ différents/même grand-parent}} = ½$, mais dans ce cas ils ne peuvent être identiques par ascendance qu'avec la probabilité de F_{t-2}, le coefficient de consanguinité des grands-parents de l'individu concerné. Par conséquent, la probabilité que deux gènes d'un individu de la génération t soient identiques par descendance et proviennent d'un même grand-parent sera de :

$$P_{\text{Id/même grand-parent}} = P_{\text{même grand-parent}} * [P_{2 \text{ mêmes/même grand-parent}} + P_{2 \text{ différents/même grand-parent}} * F_{t-2}]$$

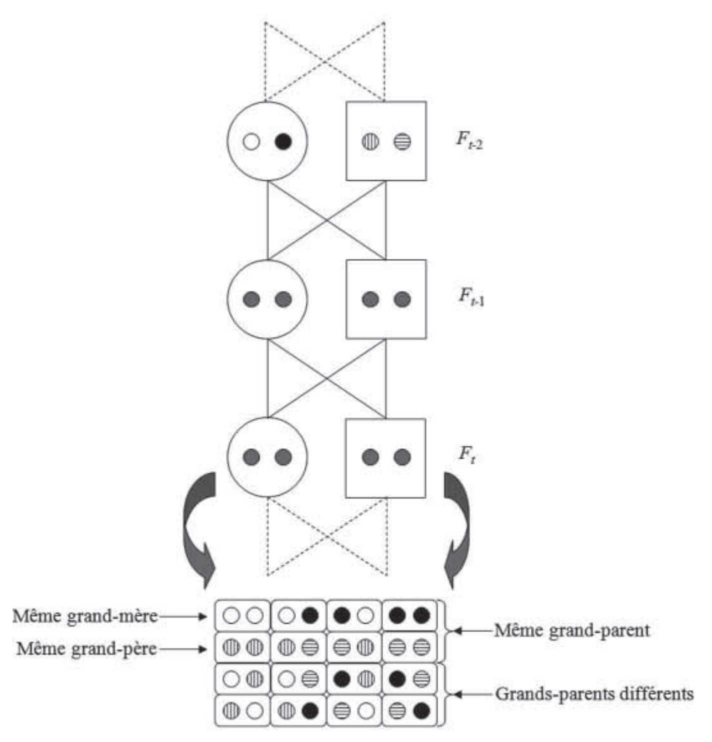


Figure 99
Évolution de la consanguinité dans un système de croisements frères-sœurs. Les femelles sont représentées par des ronds et les mâles par des carrés. Les gènes examinés sont représentés par des petits ronds. En bas sont représentés les différents petits-enfants possibles en fonction des gènes présents chez leurs deux grands-parents.

ce qui donne :

$$P_{\text{Id/même grand-parent}} = 1/2[1/2 + 1/2F_{t\text{-}2}]$$

Pour le cas où ces gènes proviennent chacun d'un grand-parent différent, ces gènes ne peuvent être identiques par ascendance que si les deux grands-parents sont apparentés.

Sachant que la probabilité de tirer deux gènes identiques par ascendance chez les deux grands-parents est égale à l'apparentement entre ces deux grands-parents et correspond très exactement à la consanguinité de leurs descendants (t - 1) $F_{t\text{-}1}$ on obtient alors :

$$P_{\text{Id/grands-parents différents}} = P_{\text{grands-parents différents}} \times F_{t\text{-}1} = 1/2F_{t\text{-}1}$$

À partir de là, il est facile de poser :

$$F_t = P_{\text{Id/même grand-parent}} + P_{\text{Id/grands-parents différents}} = 1/2[1/2 + 1/2F_{t\text{-}2}] + 1/2F_{t\text{-}1}$$

Soit, de manière plus compacte :

$$F_t = 1/4[1 + 2F_{t\text{-}1} + F_{t\text{-}2}]$$

Si on suppose alors que la proportion de croisements frère-sœur est de b et celle de croisements pangamiques de $(1 - b)$, dans une grande population avec un nombre infini d'allèles nous pouvons poser :

$$F_t = b1/4[1 + 2F_{t-1} + F_{t-2}] + (1 - b)0$$

À l'équilibre génotypique entre croisements frère-sœur et pangamie ($F_t = F_{t-1} = F_{t-2}$), nous pouvons alors poser que le F_{IS} est une mesure de la consanguinité ainsi créée et que ce dernier vérifie l'égalité :

$$F_{IS} = b\frac{1 + 2F_{IS} + F_{IS}}{4}$$

D'où on tire facilement :

$$b = \frac{4F_{IS}}{1 + 3F_{IS}}$$

Notons qu'il s'agit là d'une approximation très grossière.

Réponse 12 : Le critère d'information d'Akaike pour choisir le meilleur modèle de régression

L'AIC (*Akaike Information Criterion*) dont la valeur doit être minimale, est une mesure de la qualité d'ajustement d'un modèle statistique considéré estimé par rapport à des données. Il prend ses racines du principe d'entropie en offrant une mesure relative de la perte d'information lorsqu'un modèle est utilisé pour décrire des données réelles. On peut aussi dire qu'il correspond à un compromis entre biais et variance ou encore entre la complexité et la précision du modèle. Il n'existe pas d'AIC seuil en deçà duquel un modèle est rejeté. Il ne s'agit donc pas d'un test, mais d'un outil d'aide à la sélection du modèle le plus simple permettant d'expliquer au mieux les données, le modèle doté du plus petit AIC étant le meilleur.

Si on pose que :

$$RSS = \sum_{i=1}^{N} \hat{\varepsilon}_i^2$$

est la somme des carrés des résidus (part de la dispersion des points non expliquée par le modèle) pour un échantillon de taille N, alors on peut écrire que :

$$AIC = 2k + N[\text{Ln}(2\pi RSS/N) + 1]$$

où k est le nombre de paramètres dans le modèle.

On voit bien qu'augmenter le nombre de paramètres, même s'il permet un meilleur ajustement aux données (en diminuant RSS), augmente par ailleurs la valeur de AIC (k augmente).

Réponse 13 : La famille « quasi » des modèle linéaires généralisés

L'estimation dite *quasi-likelihood* permet de procéder à une régression sans connaître entièrement la distribution des résidus de la variable à expliquer, il faut spécifier le

« lien » (binomial ou poisson) et l'estimation se fera en tenant compte de la relation entre variance et moyenne, soit pour un lien de type binomial :

$$Var(p) = \phi \frac{p}{1-p}$$

où p est la probabilité moyenne, $Var(p)$ est sa variance et ϕ le coefficient de dispersion.

Pour plus de précisions sur les modèles *quasi-likelihood*, l'aide en ligne de R conseille les ouvrages de Cox et Snell (1981), Dobson (1983), McCullagh et Nelder (1989) (le plus souvent cité par les spécialistes) et Chambers et Hastie (1992) (très souvent cité également).

Selon mon expérience personnelle, pour les modèles logistiques, les estimations quasi peuvent conduire à des résultats aberrants, en particulier quand les occurrences d'un événement sont rares, notamment au niveau des tests (P-value = 0 alors que l'on se situe en limite de puissance dans ce cas de figure).

Réponse 14 : Calculs d'apparentement dans une population de consanguinité F (F se note aussi Q_I)

L'apparentement R entre deux individus correspond à la proportion de cas où ces deux individus portent au moins un gène identique par ascendance. R est donc égal au double de la parenté (notée Q_S) qui est la probabilité de tirer deux allèles identiques par ascendance entre deux individus de la même population. Si le système de reproduction explique la totalité de F, alors $F = F_{IS}$. La parenté entre un frère et une sœur de mêmes parents (pleins frères) est de 0,25 (1/2×1/2 = 1/4) dans une population non consanguine. Elle est de $Q_S = 0{,}25 \times (1 + F)$ dans une population de consanguinité F. Donc, dans une population où le système de croisements explique entièrement le F_{IS}, la parenté frère-sœur est de $Q_S = 0{,}25 \times (1 + F_{IS})$ et donc leur apparentement $R = 2 \times 0{,}25 \times (1 + F_{IS}) = 0{,}5 \times (1 + F_{IS})$, alors que l'apparentement entre deux individus quelconque de cette population sera de $2F_{IS}$.

Réponse 15 : Calcul du F_{IS} moyen dans des fratries d'une espèce gonochorique

Nous allons considérer un modèle IAM de mutation (beaucoup d'allèles) dans une grande population. Deux cas sont possibles si on considère que l'on a pangamie.

Soit la mère de la fratrie est homozygote *ii* avec la probabilité ~ p_i^2, soit elle est hétérozygote *ij* avec la probabilité ~ $2p_ip_j$ (p_i étant la fréquence de i dans la population) (on indique "~" car en dioecie, on approche cette valeur pour de grandes populations seulement). Dans sa descendance, la femelle homozygote produira des hétérozygotes *ij* si elle reçoit du j avec la probabilité $1\text{-}p_i$. La proportion d'hétérozygotes observés dans ce type de fratries sera donc en moyenne de :

$$H_{obs\,/\,ii} = \sum_i p_i^2 (1 - p_i)$$

La probabilité de fabriquer des *ii* dans cette fratrie est égale à p_i ou probabilité que la femelle reçoive un spermatozoïde i. La proportion attendue d'hétérozygotes sous

panmixie dans ce type de fratries sera de $2p_{i/ii}(1 - p_{i/ii})$, soit, si les *ii* sont en fréquence p_i et les *ii* en fréquence $1 - p_i$ dans la fratrie, alors $p_{i/ii} = p_i + 1/2(1 - p_i) = 1/2(p_i + 1)$ et donc $1/2(1 + p_i)(1 - p_i)$ hétérozygotes attendus dans ce type de fratrie. On attend donc dans la population :

$$H_{\exp/\,ii} = \sum_i p_i^2 \frac{1}{2}(1 + p_i)(1 - p_i)$$

$$H_{\exp/\,ii} = \frac{1}{2}\sum_i p_i^2 \left(1 - p_i^2\right)$$

Si la mère est hétérozygote *ij* avec la probabilité $2p_ip_j$, elle produit ½ de *ij* si elle reçoit du *i* ou du *j* avec la probabilité p_i et p_j respectivement et d'autres hétérozygotes si elle reçoit d'autres allèles avec la probabilité $1 - p_i - p_j$ (voir le tableau).

Mère	***i***	***j***
Père	**1/2**	**1/2**
i p_i	*ii* $1/2p_i$	*ij* $1/2p_i$
j p_j	*ij* $1/2p_j$	*jj* $1/2p_j$
autre $1-p_i-p_j$	autre hétérozygote $½(1 - p_i - p_j)$	autre hétérozygote $½(1 - p_i - p_j)$

En tout, nous obtenons dans ce type de fratrie $1/2p_i + 1/2p_j + 1 - p_i - p_j$, soit $1 - (p_i + p_j)/2$ hétérozygotes. Sur l'ensemble, nous obtenons la moyenne pondérée :

$$H_{obs\ /\,ij} = \sum_{i,\,j \neq i} 2 p_i p_j \left(1 - \frac{p_i + p_j}{2}\right)$$

La fréquence de *i* est égale à la fréquence des homozygote *ii* plus ½ de celle des hétérozygotes contenant cet allèle dans ce type de fratries. Donc $1/2p_i + 1/2[1/2p_i + 1/2p_j + ½(1 - p_i - p_j)]$, soit $½(p_i + 1/2)$. De la même façon, la fréquence de *j* sera de $½(p_j + 1/2)$ et enfin celle des autres allèles, tous hétérozygotes, sera de $½(1 - p_i - p_j)$. Par conséquent, on attend comme hétérozygotes, sous l'hypothèse de panmixie :

ij en fréquence $2½(p_i + 1/2)½(p_j + 1/2)$

i-autre en fréquence $2½(p_i + 1/2)½(1 - p_i - p_j)$

j-autre en fréquence $2½(p_j + 1/2)½(1 - p_i - p_j)$

autre-autre en fréquence

$$\sum_{k \neq i,\,j} 2\frac{1}{K-2}\frac{1}{2}\left(1 - p_i - p_j\right)\left[1 - \frac{1}{K-2}\frac{1}{2}\left(1 - p_i - p_j\right)\right]$$

où K est le nombre d'allèles que l'on suppose assez grand ici pour simplifier les choses.

Cela donne donc en moyenne pondérée :

$$H_{\exp/ij}=\sum_{i,\, j\neq i} 2p_i p_j\left[\frac{1}{2}\left(p_i+\frac{1}{2}\right)\left(p_j+\frac{1}{2}\right)+\frac{1}{2}\left(1-p_i-p_j\right)\left(1+p_i+p_j\right)+\varepsilon\right]$$

où ε est une quantité négligeable. Donc :

$$H_{\exp/ij}=\sum_{i,\, j\neq i} p_i p_j\left\{\left(p_i+\frac{1}{2}\right)\left(p_j+\frac{1}{2}\right)+\left[1-\left(p_i+p_j\right)\right]\left[1+\left(p_i+p_j\right)\right]\right\}$$

que l'on peut écrire :

$$H_{\exp/ij}=\sum_{i,\, j\neq i} p_i p_j\left[1+\left(p_i+\frac{1}{2}\right)\left(p_j+\frac{1}{2}\right)-\left(p_i+p_j\right)^2\right]$$

Par conséquent, le F_{IS} moyen attendu dans les fratries, est :

$$F_{IS/Fraterie}=1-\frac{H_{obs/ii}+H_{obs/ij}}{H_{\exp/ii}+H_{\exp/ij}}$$

En fonction des fréquences d'allèles cela donne :

$$F_{IS/Fraterie}=1-\frac{\sum_i p_i^2\left(1-p_i\right)+\sum_{i,\, j\neq i} 2p_i p_j\left(1-\frac{p_i+p_j}{2}\right)}{\frac{1}{2}\sum_i p_i^2\left(1-p_i^2\right)+\sum_{i,\, j\neq i} p_i p_j\left[1+\left(p_i+\frac{1}{2}\right)\left(p_j+\frac{1}{2}\right)-\left(p_i+p_j\right)^2\right]}$$

CQFD même si c'est plutôt moche.

Glossaire

ADN
Acide désoxyribonucléique, molécule de base de l'hérédité. En anglais DNA. Pour plus de détails, consulter n'importe quel manuel de biochimie.

AIC3
Akaike Information Criterion, de son auteur Hirotsugu Akaike (Akaike, 1974), est une mesure de la qualité d'ajustement d'un modèle statistique estimé par rapport aux données. Sa valeur dépend à la fois du nombre de paramètres du modèle et de la dispersion des données autour des valeurs attendues du modèle. Le meilleur modèle est celui qui présente le plus petit *AIC*. Plus de détails sont donnés en réponse 12.

Allèle
État héréditaire dans lequel un locus se présente. Chez les diploïdes, chaque individu présente deux allèles à chaque locus. Ces allèles peuvent être identiques (homozygote) ou différents (hétérozygote).

***Allelic dropout* (pas de traduction simple)**
Phénomène qui fait qu'un allèle n'est pas vu en face d'un autre à cause, par exemple, d'une compétition pour la Taq polymérase lors d'une PCR avec peu d'ADN. Dans ce cas, un seul allèle se trouve amplifié et l'individu est erronément interprété homozygote.

Améiotique
Processus de reproduction qui se déroule sans intervention de la méiose.

Apostatique (sélection)
Processus sélectif qui avantage les génotypes ou phénotypes les plus rares. Par définition une sélection qui maintient une diversité stable.

Arithmétique
Voir Moyenne.

Assortative mating
Processus d'appariement préférentiel des partenaires sexuels qui se ressemblent le plus phénotypiquement (voir aussi homogamie).

ARN
Acide ribonucléique, normalement transcrit de l'ADN et ensuite traduit en protéine.

Autoincompatibilité
Système interdisant l'autofécondation.

Autosome
Désigne un chromosome ordinaire présent en paire dans chaque zygote ou individu diploïde normal (antonymique de hétérosome).

Auto-stop
Hitchhiking en anglais. Phénomène sélectif au cours duquel la sélection à un locus entraîne des modifications de la distribution des fréquences génotypiques à un autre locus lié physiquement (proche sur le même chromosome) ou statistiquement quand le mode de reproduction est fermé (clonalité, autofécondation…).

Bottleneck
En français goulot d'étranglement. Désigne un processus démographique durant lequel une population subit une chute brutale d'effectif (nombre d'individus reproducteurs).

Cline
Généralement géographique, il correspond à l'augmentation ou la diminution graduelle des fréquences alléliques à un ou plusieurs loci le long d'un axe géographique et/ou d'un gradient écologique.

Clonalité
Reproduction asexuée où la descendance est produite sans subir ni ségrégation ni recombinaison (améiotique) et se retrouve donc génétiquement strictement identique à l'individu parental, à la mutation somatique près.

CMH (MHC en anglais)
Complexe majeur d'histocompatibilité. Complexe de gènes qui détermine (entre autres) la reconnaissance du soi et du non soi. Voir aussi HLA.

Coalescence
Phénomène qui décrit l'ascendance commune de deux gènes d'une population. Le temps de coalescence décrit, par exemple le nombre de générations qu'il est nécessaire de remonter pour atteindre le gène ancêtre commun de deux gènes pris au hasard dans la population étudiée.

Codominant
Décrit un marqueur génétique pour lequel tous les hétérozygotes sont distinguables des homozygotes (ni dominant, ni récessif).

Consanguinité
Indique la proportion de loci identiques par descendance au sein des individus, résultant d'un système de reproduction fermé (autofécondation, croisement entre

apparentés) ou d'une taille limitée de la population. Notons que dans le cas où cette consanguinité (probabilité d'identité par descendance intra-individuelle) ne résulte que de la taille de la population, celle-ci devient égale à l'apparentement entre individus de cette population (probabilité d'identité par descendance interindividuelle).

Crossing-over
Phénomène chromosomique intervenant lors de la méiose et consistant à un échange de portions plus ou moins grandes et en principe de mêmes tailles des chromosomes homologues, précédant la formation des gamètes et résultant en un réassortiment (ou recombinaison) intra-chromosomique.

Dème
Unité démographique d'individus appartenant à la même unité de reproduction ou partageant les mêmes paramètres de régulation démographique (par exemple, entre lesquels la compétition intra-spécifique est maximale), synonyme de sous-population.

Dérive génétique
Décrit le processus par lequel les fréquences alléliques changent d'une génération à l'autre à cause d'un échantillonnage aléatoire des individus (gamètes, zygotes, adultes) devant survivre pour participer à la reproduction de la génération suivante dans une population de taille finie.

Déséquilibre de liaison
Exprime une association non aléatoire entre différents loci (souvent pris par paire). Beaucoup de facteurs différents peuvent influencer le déséquilibre de liaison (structure de la population, système de reproduction, sélection, etc.).

Déviance
Terme utilisé en régression linéaire généralisée (GLiM) qui décrit la dispersion de résidus autour des valeurs attendues définies par le modèle. Consulter des ouvrages spécialisés pour des définitions plus strictes.

Dioïque
Synonyme de gonochorique (terme un peu désuet aujourd'hui) et signifiant que l'espèce étudiée est séparée en deux sexes (femelles et mâles) (antonymique de monoïque).

Diploïde
Caractérise un organisme ou une cellule possédant un matériel génétique (chromosomes) en double, à l'exception des chromosomes sexuels quand ces derniers existent.

Directionnelle (sélection)
Processus sélectif tendant à accroître ou décroître (une seule direction) la fréquence d'un allèle (ou d'un phénotype) dans une population.

Disruptive (sélection)
Sélection directionnelle dans chaque sous-population, mais divergente d'une sous-population à l'autre.

Dominant
Caractérise un marqueur génétique pour lequel un des allèles masque à l'état hétérozygote les autres allèles. Caractérise aussi un tel allèle (antonymique de récessif).

Dropout
Voir *Allelic dropout*.

Épistatique (par exemple sélection)
Forme de déterminisme génétique où les différentes formes d'un gène (allèles) vont avoir différentes répercussions sur l'expression phénotypique des allèles d'un autre locus. C'est typiquement le cas des gènes de régulation.

Exon
Partie d'un gène conservée lors du passage de l'ARN de transfert à l'ARN messager (épissage) et qui sera donc traduite en protéine.

Fréquence dépendante (sélection)
Voire Apostatique.

Gamète
Cellule sexuelle normalement haploïde. Chez les animaux, les gamètes femelles sont appelés ovules et les gamètes mâles spermatozoïdes.

Gaussienne
Se dit d'une distribution de données ordinales continues en forme de cloche (voir aussi Poissonienne et Logistique).

Gène
Une portion d'ADN qui code pour une fonction, c'est-à-dire transcrite en ARN de transfert et ensuite en ARN messager (ou mRNA). L'ARN messager devant lui-même être traduit en molécule active tel un enzyme.

Génotype
Donne la composition allélique complète d'un individu à un locus donné ou à une série de loci spécifiques (quand précisé).

Géométrique
Voir Moyenne.

Germinal
Qui provient de la lignée du même nom, cellules souches des cellules sexuelles (ou gamètes).

Gonochorique
Terme un peu désuet aujourd'hui synonyme de dioïque (antonymique d'hermaphrodite).

Goulot d'étranglement
voir *Bottleneck*.

Haploïde
Caractérise un organisme ou une cellule avec un matériel génétique (chromosomes) présent en un seul exemplaire. Les cellules sexuelles (gamètes) sont typiquement haploïdes.

Harmonique
Voir Moyenne.

Hermaphrodite
Se dit d'une espèce à reproduction sexuée où chaque individu peut assurer les deux fonctions femelle et mâle (antonymique de gonochorique).

Hétérogamie
Processus de reproduction sexuée au cours duquel les individus ou leurs gamètes sont d'autant plus attirés l'un par l'autre (pour la reproduction) qu'ils diffèrent génétiquement (antonymique de l'homogamie).

Hétérosis
Phénomène sélectif concernant l'ensemble du génome au cours duquel les individus les plus hétérozygotes (en nombre de loci) sont favorisés (survie et/ou reproduction accrues).

Hétérosome
Synonyme de chromosome sexuel. Chez les espèces dioïques, le déterminisme du sexe peut être chromosomique. Dans ce cas, la composition en chromosome sexuel diffère entre les deux sexes (chromosomes XY des mammifères, chromosomes ZW des oiseaux) (antonymique d'autosome).

Hétérozygote
État d'un locus chez un individu diploïde présentant deux allèles différents (antonymique d'homozygote).

Hitchhiking
Voir Auto-stop.

HLA
Human Leukocyte Antigen, équivalent du MHC des vertébrés pour l'homme.

Homogamie
Processus de reproduction sexuée au cours duquel les individus ou leurs gamètes sont d'autant plus attirés entre eux (pour la reproduction) qu'ils se ressemblent génétiquement (antonymique de l'hétérogamie, voir aussi *assortative mating*).

Homoplasie
Phénomène décrivant l'identité entre deux allèles ne résultant pas d'une parenté commune récente, qui sont alors qualifiés d'identiques par état. Les microsatellites, et plus encore les SNP, sont par nature homoplasiques.

Homozygote
État d'un locus chez un individu diploïde présentant deux fois le même allèle (antonymique d'hétérozygote).

IAM (*Infinite Allele Model*)
Modèle de mutation où chaque mutation génère un nouvel allèle qui n'existait pas auparavant dans la population, et qui sera définitivement perdu s'il disparaît. Ne permet aucune homoplasie.

Îles (modèle en)
Modèle théorique de population structurée en n dèmes de tailles identiques N composés à chaque génération non chevauchante de $(1-m)N$ individus autochtones et de mN individus migrants provenant aléatoirement de n'importe quel des n dèmes.

Inbreeding
Voir Consanguinité.

Infinite island model
Ou modèle en îles infini. Modèle en îles avec un nombre infini de sous-populations.

Infra-population
Utilisé en parasitologie pour désigner l'ensemble des individus de la même espèce de parasite contenus dans un individu hôte.

Intron
Partie d'un gène qui ne sera pas traduite en protéine, car éliminée lors du passage de l'ARN de transfert vers l'ARN messager (phénomène d'épissage) (antonymique d'exon).

Island model
Modèle en îles. Un modèle théorique de population subdivisée, à générations non chevauchantes, où les individus sont distribués dans *n* sous-populations (îles) chacune de taille *N* et construite à chaque génération de *mN* migrants en provenance de n'importe laquelle de *n* sous-populations et (1-*m*)*N* résidents ou individus philopatriques.

KAM (*K Allele Model*)
Modèle de mutation en nombre fini (*K*) d'allèles. Modèle de mutation où chaque mutation change un allèle dans un autre allèle parmi les *K* existants, avec la même probabilité, y compris en lui-même. Plus *K* est petit, plus fréquente est l'homoplasie.

Linkage disequilibrium
Voir Déséquilibre de liaison.

Locus
Décrit une portion de l'ADN située dans une position spécifique du génome. Un locus ne correspond pas nécessairement à un gène.

Logistique
Se dit d'une distribution de données disjointes en vrai et faux (ou 0 et 1) (voir aussi Gaussienne et Poissonienne).

Métapopulation
Une population composée de plusieurs unités (sous-populations ou dèmes). Chaque sous-population peut être caractérisée par une probabilité d'extinction ou de recolonisation. Les dèmes peuvent aussi être stables (comme dans un modèle en îles).

Méiose
Processus de production des cellules de la reproduction sexuée ou gamètes. C'est au cours de ce processus qu'ont lieu la ségrégation des allèles à chaque locus et la recombinaison entre loci, pour aboutir à la formation de cellules haploïdes.

Microsatellite
Élément constitutif de l'ADN des eukaryotes. Il s'agit de courtes séquences répétées d'ADN réparties dans le génome et, la plupart du temps, sans fonction connue.

Mutation
Erreur héréditaire intervenant lors de la duplication de l'ADN.

Monoïque
Synonyme d'hermaphrodite (antonymique de dioïque).

Moyenne
Valeur unique x que devraient avoir les N individus i d'une population (ou d'un échantillon) pour que leur total soit inchangé. Il en existe trois types la moyenne arithmétique (la plus courante) $\overline{x_{Ari}} = \frac{1}{N}\sum_{i=1}^{N} x_i$; la moyenne géométrique $\overline{x_{Geo}} = \sqrt[N]{\Pi_{i=1}^{N} x_i}$ ou racine Nième des N produits $x_1 \times x_2 \times \ldots \times x_i$; la moyenne harmonique $\overline{x_{Har}} = \frac{1}{\sum_{i=1}^{N} \frac{1}{x_i}}$.

Neighbourhood model
Modèle en voisinage. Un modèle théorique de population structurée où la migration de chaque individu est limitée par la distance, de telle sorte que l'apparentement entre individus devient une fonction décroissante de la distance qui les sépare, même en l'absence de toute barrière ou délimitation visible.

Neutre
Définit un locus ou un caractère dont le polymorphisme n'est soumis à aucune pression sélective d'aucune sorte (antonymique de sélectionné).

Ordinales
Qualifie des données que l'on peut ordonner (comptages ou mesures).

Overdominance
Superdominance. Processus sélectif au cours duquel la survie et/ou le succès reproducteur d'un individu se trouve augmentés si cet individu est hétérozygote à un locus donné.

Ovule
Gamète femelle.

Pangamie
Décrit un mode d'accouplement aléatoire (indépendant du génotype) des individus d'une population à reproduction sexuée.

Panmixie
Décrit un mode de reproduction sexuée où les zygotes sont formés par rencontre aléatoire de tous les gamètes de la population.

Parthénogenèse
Du grec παρθενος (partenos = vierge) and γένεσις (genèse), quand une mère produit des filles à partir d'ovules non fécondés.

Pas japonais (Modèle en)
Stepping-stone model. Modèle théorique de population subdivisée où les migrants ne s'échangent qu'entre sous-populations adjacentes.

PCR
Polymerase Chain Reaction, qui permet d'amplifier une portion d'ADN encadrée de séquences connues à partir de deux amorces d'ADN courtes spécifiques d'une zone de ces séquences flanquantes (plus de précisions dans Google).

Phénotype
Il s'agit de l'expression d'un caractère éventuellement héréditaire (comme la couleur des yeux). Pour des marqueurs codominants, le phénotype peut directement être traduit en génotype.

Philopatrique
Se dit d'un individu qui montre une tendance significative au retour vers son lieu de naissance.

Phylogéographie
Discipline visant à établir les relations de « parenté » entre populations géographiquement éloignées de la même espèce afin, par exemple, d'établir un scénario de colonisation de l'aire géographique occupée par cette espèce.

Pléïotropique
Se dit d'une sélection ou de l'effet d'un seul gène (ou famille de gènes) qui affecte deux caractères différents, comme par exemple les gènes du CMH (HLA chez l'homme) qui affectent à la fois le système immunitaire et la sélection du partenaire sexuel.

Poissonienne
Se dit d'une distribution de données ordinales discontinues (comptages) suivant une courbe en cloche (voir gaussienne et logistique).

Polymorphe
Condition qui décrit qu'un locus est variable d'un individu à l'autre, c'est-à-dire qu'il présente plus d'un allèle dans l'échantillon d'individus génotypés.

Population
Groupe d'individus partageant les mêmes paramètres démographiques, en particulier la régulation de la population, et partageant une ascendance commune plus probable avec les individus de la même unité qu'avec des individus d'autres populations définies comme telles, exception faite des migrants, bien entendu.

Purine
Base, constituant essentiel des nucléotides eux-mêmes éléments de base des acides nucléiques (ARN et ADN), complémentaires des Pyrimidines. Il en existe deux :

l'adénine (A) complémentaire de la thymine (T dans l'ADN) et de l'uracile (U dans l'ARN) et la guanine (G) complémentaire de la cytosine (C).

Pyrimidines
Base, constituant essentiel des nucléotides eux-mêmes éléments de base des acides nucléiques (ARN et ADN), complémentaires des purines. Il en existe trois : la thymine (T), l'uracile (U qui prend la place de T dans l'ARN) et la cytosine (C).

Récessif
Caractérise un allèle qui est masqué quand hétérozygote avec un autre allèle (antonymique de dominant).

Recombinaison
Processus durant lequel les allèles de loci différents, auparavant associés, se retrouvent dissociés et réassociés à d'autres allèles. C'est ce qui se passe durant la méiose entre loci de chromosomes différents ou du même chromosome après crossing-over.

Ségrégation
Processus intervenant lors de la méiose et durant lequel les deux allèles de chaque locus se trouvent séparés pour devenir indépendants (dans des gamètes différents).

Sélection
Processus durant lequel la survie et/ou le succès reproducteur d'un individu dépend de son phénotype ou de son génotype d'une manière plus ou moins directe.

Sélectionné
S'applique pour un locus ou un caractère soumis à sélection (antonymique de neutre).

Selfing
Voir autofécondation.

Sex-ratio
Ratio du nombre de mâles sur le nombre de femelles dans une population. Égal à un quand il est équilibré.

SMM (*Stepwise Mutation Model*)
Mécanisme de mutation au cours duquel chaque mutation augmente ou diminue, avec une égale probabilité, la taille de l'allèle d'une unité (*step*) pré-définie. Ce mode de mutation génère beaucoup d'homoplasie et aboutit au fait qu'une ressemblance de taille peut se traduire par une ascendance commune de deux allèles.

SNP
Single nucleotide polymorphism. Marqueurs génétiques déterminés par la mutation d'un site (paire de base) de l'ADN, avec en général deux allèles possibles, car les transitions sont beaucoup plus fréquentes que les transversions.

Somatique
Ce qui vient du soma, c'est-à-dire n'impliquant pas les cellules de la lignée dite germinale (antonymique de germinal).

Sous-dominance
Processus sélectif au cours duquel les individus hétérozygotes à un locus donné montrent une survie et/ou un succès reproducteur réduit.

Sous-population
Voir Dème.

Spermatozoïde
Gamète mâle.

Stepping-stone model
Voir Pas japonais.

Superdominance
Voir *Overdominance*.

Taq polymérase
Enzyme : DNA polymérase extraite de l'extrémophile *Thermophilus aquaticus* capable de synthétiser de l'ADN à très hautes températures et utilisée pour les réactions de PCR.

Tore
Définit la surface d'une figure géométrique en trois dimensions ayant la forme d'une bouée ou d'un donut (pour les plus gourmands).

TPM (*Two Phase Model*)
Modèle de mutation combinant le KAM et le SMM avec une proportion variable de mutations générées par l'un ou l'autre des mécanismes correspondants.

Transition
Mutation ponctuelle consistant au remplacement d'une purine par une autre purine (A<=>G) ou d'une pyrimidine par une autre pyrimidine (C<=>T) (antonymique de transversion).

Transversion
Mutation ponctuelle consistant au remplacement d’une purine par une pyrimidine ou d’une pyrimidine par une purine (A<=>T, A<=>C, G<=>C, G<=>T) (antonymique de transition).

Underdominance
Voir Sous-dominance.

Végétative
Mode de reproduction purement asexuée où un individu donne naissance à plusieurs autres individus par simple division (mitose ou scissiparité).

Vigueur hybride
Voir Hétérosis.

Voisinage (Modèle en)
Voir *Neighbourhood model*.

Wahlund (Effet)
Diminution de l’hétérozygotie observée que produit le mélange dans un même échantillon d’individus hétérogènes génétiquement.

Zygote
Résultat de la fusion de deux gamètes. Le terme œuf est aussi parfois usité.



Annexe

Tableau 1
Liste des logiciels cités et/ou utilisés, leur lien pour téléchargement et références bibliographiques quand elles existent.

Logiciel	URL	Auteur(s) et date	Références
BAPS	http://www.rni.helsinki.fi/~jic/bapspage.html	Corander et Marttinen (2005)	Corander *et al.* (2003, 2004).
Bottleneck	http://www.montpellier.inra.fr/URLB/bottleneck/bottleneck.html	Piry *et al.* (1999)	Cornuet et Luikart (1996)
CREATE	http://www.lsc.usgs.gov/CAFL/Ecology/Software.html	Coombs *et al.* (2008)	Coombs *et al.* (2008)
Estim	http://www.ecoanthropologie.cnrs.fr/spip.php?article296	Vitalis et Couvet (2001)	Vitalis et Couvet (2001a)
Flock	http://www.bio.ulaval.ca/no_cache/departement/professeurs/fiche_des_professeurs/professeur/11/13/	Duchesne *et al.* (2009)	Duchesne et Turgeon (2009)
Fstat	http://www2.unil.ch/popgen/softwares/fstat.htm	Goudet (2003)	Goudet (1995)
Genepop 3.4	http://wbiomed.curtin.edu.au/genepop/ ou http://ftp.cefe.cnrs.fr/PC/MSDOS/GENEPOP/	Raymond et Rousset (2003)	Raymond et Rousset (1995)
Genepop 4	http://kimura.univ-montp2.fr/~rousset/Genepop.htm	Rousset (2008)	Rousset (2008)
Genepop 4 on the web	http://genepop.curtin.edu.au/	Rousset (2008)	Rousset (2008)
Genetix	http://www.univ-montp2.fr/~genetix/genetix/genetix.htm	Belkhir *et al.* (2004)	Non publié
HierFstat	http://www2.unil.ch/popgen/softwares/hierfstat.htm	Goudet (2005)	Goudet (2005)
INSTRUCT	http://cbsuapps.tc.cornell.edu/InStruct.aspx	Gao *et al.* (2007)	Gao *et al.* (2007)
MEGA	http://www.megasoftware.net/	Kumar *et al.* (2005)	Kumar *et al.* (2004)

Tableau 1 (suite)

Logiciel	URL	Auteur(s) et date	Références
Micro-Checker	http://www.microchecker.hull.ac.uk/	Van Oosterhout *et al.* (2004)	Van Oosterhout *et al.* (2004)
MLNE	http://www.zoo.cam.ac.uk/ioz/software.htm	Wang et Whitlock (2003)	Wang et Whitlock (2003)
ML-RELATE	http://www.montana.edu/kalinowski/Software.htm	Kalinowski (2008)	Kalinowski *et al.* (2006)
MSA	http://i122server.vu-wien.ac.at	Diringer et Schlötterer (2002)	Diringer et Schlötterer (2002)
Multilocus	http://www.bio.ic.ac.uk/evolve/software/multilocus	Agapow et Burt (2000)	Agapow et Burt (2001)
MultiTest V.1.2	http://gemi.mpl.ird.fr/SiteSGASS/SiteTDM/Programs	De Meeûs *et al.* (2009)	De Meeûs *et al.* (2009)
NeEstimator	http://www.dpi.qld.gov.au/fishweb/11629.html	Peel *et al.* (2004)	Non publié
PCAGen	http://www2.unil.ch/popgen/softwares/pcagen.htm	Goudet (1999)	Non publié
PGD-Spider	http://cmpg.unibe.ch/software/PGDSpider/	Lischer et Excoffier (2012)	Lischer et Excoffier (2012)
R	http://www.R-project.org	R Development Core Team (2004)	Non publié
RMES	http://www.cefe.cnrs.fr/genetique-et-ecologie-evolutive/patrice-david	David *et al.* (2007)	David *et al.* (2007)
SGM	jerome.goudet@unil.ch	Goudet (1999)	Goudet (1999)
Structure	http://pritch.bsd.uchicago.edu/software.html	Pritchard *et al.* (2002) ; Falush *et al.* (2003)	Pritchard *et al.* (2000)

Tableau 2
Les différents types d'analyses traitées dans ce manuel, description théorique telle que décrite dans la partie I (Théorie), logiciel utilisé et application dans la partie II (Exemple). Excel indique que le test peut être effectué dans une feuille de calcul de type Excel.

Analyse		Théorie	Logiciels[1]	Exemples
Déséquilibre de liaison		p 65, 80	Fstat	p 118, 195, 220, 225, 256
			Genetix	
			Multilocus	
Identifier les tests significatifs : ajuster une série de tests au seuil Bonferroni		p 82	Excel	p 120, 147, 182, 195
Hardy-Weinberg dans les sous-échantillons		p 76	Fstat	p 118, 195, 219, 225, 256
			Genepop	
			Genetix	
Calcul d'intervalles de confiance des *F*		p 70	Fstat	p 120, 223, 235, 256
Allèles nuls et *stuttering*		p 105, 111	Micro-Checker	p 125, 196
Allèles nuls et régression		p 107	R	p 165, 197, 222
Allèles nuls et autofécondation		p 107	RMES	
Ajuster à une proportion attendue / test binomial exact		p 83	R	p 128, 148, 174, 177, 195, 222, 226
Dominance des allèles courts		p 107	R	p 129, 167
Structure cachée (effet Wahlund)	« Panmixie »	p 98	BAPS	p 135, 199, 234
			Flock	p 234
	Avec autofécondation		INSTRUC	
Test de Wilcoxon pour données appariées			R	p 143, 186, 197
NJTree sur distances génétiques		p 98	MSA et MEGA	p 144, 254, 283
Structure génétique sexe-spécifique		p 91	Fstat	p 147, 150, 226
Combiner des tests indépendants	Binomial généralisé	p 83	MultiTest	p 148, 181, 203, 222, 233, 243
	Z de Stouffer		Excel	p 181
F hiérarchiques		p 62	HierFstat (R)	p 153, 156, 210, 223
Isolement par la distance	Entre populations : Matrices	p 88	Genepop 3	p 157
	Entre populations : GPS		Genepop 4	p 236
	Entre individus : GPS		Genepop 4	p 200

Tableau 2 (suite)

Analyse		**Théorie**	**Logiciels**[1]	**Exemples**
Test de Mantel	Matrices carrées GPS Matrices rectangles	p 88	Genepop 3 Genepop 4 Fstat	p 157 p 200, 236 p 228
Effectifs efficaces	LD Bartley Corrélations intra et inter loci LD Waples & Do F_{IS} Balloux Goulot d'étranglement Temporel Waples Spatio-temporel Wang et Whitlock	p 101	NeEstimator Estim LDNe Fstat ou Genetix et Excel Bottleneck NeEstimator MLNe	p 160, 203 p 203, 239 p 204, 239 p 203, 239 p 220, 244 p 261 p 261
Densités efficaces et dispersion		p 88	Excel	p 165, 208, 241
Régression logistique Régression quasi-binomiale			R	p 170,
Test exact de co-occurence			Fstat ; Genetix	p 174
Test planifié de différenciation par paires de sous-échantillons			Fstat	p 179, 182, 253, 258
Structure génétique pathogène spécifique		p 91	Fstat	p 182
Conversion de données au bon format		p 70	Create PGD-Spider	p 191, 196, 220
Test de corrélation de Spearman			R	p 198
Simulations de populations structurées		p 29-65	Easypop	p 213
Différenciation globale		p 43, 78	Fstat	p 226
Pangamie		p 77	ML Relate et Fstat	p 228
Nombre de migrants, *Nm*	Isolement par la distance 2D Modèle en île Spatio-temporel Wang et Whitlock	p 90 p 49 p 101	Genepop 3 ou 4 et Excel Fstat et Excel MLNe	p 160 p 261 p 267
Taille clonale	Nombre infini de dèmes Deux dèmes Un dème isolé	p 40	Papier, crayon et gomme	p 268, 272 p 268, 275 p 268, 279

[1] Logiciels proposés, mais il en existe d'autres qui peuvent être utilisés.
LD *(linkage disequilibrium)* : déséquilibres de liaison. 2D : en deux dimensions

Table des matières



Imprimé en France. - JOUVE, 1, rue du Docteur Sauvé, 53100 MAYENNE
N° 2041831D. - Dépôt légal : décembre 2012

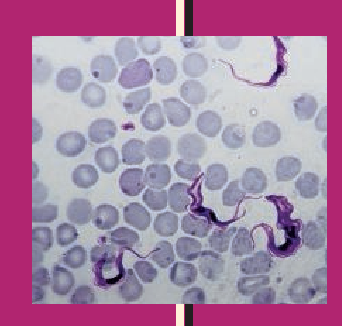


La compréhension de l'épidémiologie d'une maladie infectieuse ou parasitaire passe par une connaissance minimale du fonctionnement des populations vivantes concernées. Ainsi, pour remédier à la difficulté d'obtenir des observations directes sur la biologie des populations naturelles, notamment dans les pays du Sud, l'utilisation de marqueurs génétiques permet d'avoir accès, à travers des méthodes indirectes, à des informations clés sur la biologie des agents pathogènes et de leurs vecteurs : écologie, mode de reproduction, déplacements, taille des populations, etc.

Un outil précieux dans le domaine de la santé, où l'analyse de la variation spatio-temporelle des marqueurs génétiques peut ainsi être utilisée pour caractériser la dynamique des populations de parasites et de leurs vecteurs, pour connaître l'évolution d'une maladie infectieuse ou parasitaire, évaluer les risques d'invasions ou d'épidémie, le potentiel de diffusion de gènes résistants, anticiper les stratégies de lutte…

Ce manuel didactique présente les principales méthodes de la génétique des populations naturelles et les modèles de base utilisés pour les inférences, avec des cas concrets d'applications à destination des étudiants et personnels de santé. Plusieurs jeux de données sont analysés pas à pas dans un CD-ROM qui accompagne l'ouvrage.

Thierry De Meeûs est chercheur à l'IRD, spécialisé en écologie évolutive et en génétique des populations dans les systèmes hôte-parasite-vecteur (laboratoire Intertryp, UMR 177 IRD/Cirad). Il mène en parallèle une activité d'enseignement et de recherche et travaille actuellement au Cirdes (Burkina Faso) sur les trypanosomoses africaines et les mouches tsétsé. Il est co-auteur de nombreuses publications dans différentes revues ou ouvrages collectifs.



35 €

ISBN 978-2-7099-1732-2
ISSN 1142-2580



IRD
44, bd de Dunkerque
13572 Marseille cedex 02
editions@ird.fr
www.editions.ird.fr